

G

Gas Expanded Liquids for Sustainable Catalysis

BALA SUBRAMANIAM

Department of Chemical and Petroleum Engineering,
The Center for Environmentally Beneficial Catalysis,
University of Kansas, Lawrence, KS, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Thermodynamic and Physical Properties of GXLs

Catalytic Reactors for Investigating GXLs

Applications in Homogeneous Catalysis

Multiphase Catalysis

An Example of CXL-Based Process Development

Including Sustainability (Economic and Environmental Impact) Analyses

Summary and Future Directions

Acknowledgments

Bibliography

Glossary

Carbon selectivity Refers to the fraction of the carbon in a hydrocarbon-based feed that is utilized in making the desired product.

Compressible gas A gas in the vicinity of its critical temperature wherein it is highly compressible with pressure, either condensing or attaining liquid-like densities as the pressure approaches or exceeds its critical pressure. Below the critical temperature, a compressible gas will typically condense at sufficiently high pressures to produce a liquid phase.

Gas-expanded liquids When a liquid such as an organic solvent is pressurized with a compressible gas, the liquid phase will volumetrically expand if

the gas dissolves in it. The volumetrically expanded liquid phase is termed as a gas-expanded liquid. When the pressure is released, the dissolved gas will escape from the liquid phase causing the liquid phase to contract to its original volume.

Homogeneous catalysis Refers to a process wherein the reactants, catalyst, and products are soluble in a single phase and the catalytic reaction occurs in that phase.

Multiphase catalysis Refers to a process wherein the reactants, catalyst, and products are present in two or more different immiscible phases separated by phase boundary(ies). The reaction typically occurs at a boundary between two phases.

Renewable feedstock Refers to a feedstock from nature whose use has minimal adverse impact on the ecosystem and that has the ability to manifest itself again in nature in a matter of a few months to a few years rather than in hundreds to thousands of years.

Solvent engineering Exploiting the synergies between catalysis and solvent media for enhancing rates, selectivity, and separations in a sustainable manner.

Supercritical fluid A substance that is above its critical pressure (P_c) and critical temperature (T_c). For applications in catalysis and separations, the near-critical region [$0.9\text{--}1.2 T_c$ (in K) and $0.9\text{--}2 P_c$] wherein small changes in temperature and/or pressure yield relatively large changes (from gas-like to liquid-like values) in density and transport properties, is generally of interest.

Sustainability Sustainability of a catalytic process refers to the long-term environmental, social, and economic viability of the process.

Turnover frequency (TOF) Quantifies the intrinsic activity of a catalyst in converting the reactants to products. It is usually expressed in terms of the rate at which the reactants are converted to products [(moles of substrate converted)/(gram atoms of catalyst used)/(time)].

Definition of the Subject

The modern-day chemical industry relies mostly on fossil fuel (such as petroleum, natural gas, and coal)–based feedstock. There are several megaton industrial catalytic processes that produce essential commodities for everyday life but present challenges with respect to reducing environmental footprints and enhancing sustainability. Examples of such processes include the homogeneous hydroformylation of higher olefins, the selective oxidation of light olefins to their corresponding epoxides, and the oxidation of *p*-xylene to produce terephthalic acid. For a targeted product, there are several possible scenarios for developing sustainable alternatives to conventional technologies. These include (a) developing greener process technologies based on existing feedstock, (b) replacement of fossil fuel–based feedstock with renewable ones such as those derived from biomass (which will also entail the development of new chemistries and process technologies), or (c) replacement of the target product itself with alternate candidates from renewable feedstocks. This entry will discuss the potential of gas-expanded liquids (GXLs), a relatively new class of solvents, for developing alternative and more sustainable catalytic processes.

A gas-expanded liquid (GXL) phase is generated by dissolving a compressible gas such as CO₂ or a light olefin into the traditional liquid phase at mild pressures (tens of bar) [1]. When CO₂ is used as the expansion gas, the resulting liquid phase is termed a CO₂-expanded liquid or CXL. GXLs combine the advantages of compressed gases such as CO₂ and of traditional solvents in an optimal manner. GXLs retain the beneficial attributes of the conventional solvent (polarity, catalyst/reactant solubility) but with higher miscibility of permanent gases (O₂, H₂, CO, etc.) compared to organic solvents at ambient conditions and enhanced transport rates compared to liquid solvents [2–5]. The enhanced gas solubilities in GXLs have been exploited to alleviate gas starvation (often encountered in homogeneous catalysis with conventional solvents), resulting in a one to two orders of magnitude greater rates than in neat organic solvent or scCO₂. *Environmental advantages* include substantial replacement of organic solvents with environmentally benign CO₂. *Process advantages* include reduced flammability due to CO₂

presence in the vapor phase and milder process pressures (tens of bar) compared to scCO₂ (hundreds of bar). GXLs thus satisfy many of the attributes of an ideal alternative solvent.

Introduction

Solvent usage has often been linked to waste generation and associated environmental and economic burdens [6, 7]. Within the last two to three decades, many research groups have investigated benign alternate media for performing chemical reactions [8–12]. Examples of such media include supercritical CO₂ (scCO₂) [13–20], water [21–23], gas-expanded liquids (GXLs) [1, 24–27], ionic liquids (ILs) [28–30], and switchable solvents [31–35]. While scCO₂ is “generally regarded as safe,” its nonpolar nature renders it unsuitable for most homogeneous catalysis involving polar transition metal complexes. Further, scCO₂ media require operating pressures in excess of 100 bar. The use of either supercritical or near-critical water ($P_c = 220.6$ bar; $T_c = 373.9^\circ\text{C}$) requires rather harsh operating pressures and temperatures. The use of ionic liquids as tunable media for catalysis either alone or in combination with scCO₂ [36] shows much promise.

The ideal alternative solvent, in addition to being considered green, should typically satisfy the following criteria: (a) retain the beneficial aspects of the conventional solvent (polarity, catalyst/reactant solubility) being replaced, (b) facilitate facile product/catalyst separation, (c) enhance process safety, and (d) operate at mild pressures (tens of bar) for economic viability. The qualitative principles of green chemistry [37] and green engineering [38] provide valuable guidelines for developing greener process alternatives. Some of these principles include the use of renewable and abundant resources as feedstock, nonhazardous reagents as reaction and separation media, inherently safe process design, and process intensification at mild conditions. However, a reliable assessment of overall “greenness” and sustainability requires quantitative comparison with conventional processes using metrics such as atom economy, the E-factor (amount of waste produced/unit of desired product) [6], toxic emissions potential [39], and process economics. Such analyses also provide guidance in identifying potential process

improvement opportunities and establishing performance metrics for sustainability.

For a targeted product from a given feedstock, the basic elements of a catalytic process or system include the catalyst, media, multiphase reactor, and separation. Effective integration of these elements into a sustainable technology requires a multiscale approach [40]. This entry highlights reported examples of catalytic process concepts with gas-expanded liquids and multiscale approaches to develop such processes for large-scale catalytic technologies. The examples include catalytic hydrogenations, hydroformylations, and selective oxidations. Specifically, it is shown how the tunable physicochemical properties of GXLs can be effectively exploited to promote sustainable catalysis. An example of quantitative economic and environmental impact analyses is also presented to show how such an assessment validates and facilitates the design and development of sustainable systems that are also practically viable.

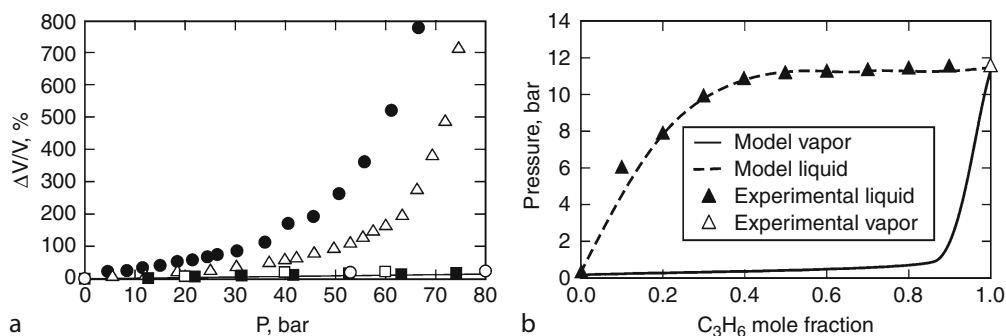
Thermodynamic and Physical Properties of GXLs

A gas in the vicinity of its critical temperature [$0.9\text{--}1.2 T_c$ (K)] is highly compressible and attains liquid-like densities when sufficiently compressed to near-critical or supercritical pressures. As an example, consider CO_2 ($P_c = 71.8$ bar; $T_c = 304.1$ K) as an expansion gas. At 40°C (313.15 K and $T/T_c = 1.03$), CO_2 is highly

compressible attaining liquid-like densities as its critical pressure is approached. In such a compressed state, CO_2 dissolves in a liquid phase and volumetrically “expands” the liquid phase forming a GXL. When the system pressure is released, the dissolved CO_2 escapes from the liquid phase contracting the liquid phase. Gases whose critical temperatures are far removed from the reaction temperature do not exhibit such compressibility and are generally incapable of expanding solvents.

Liquids expand to different extents in the presence of CO_2 pressure, depending on the ability of the liquids to dissolve CO_2 . The equipment and experimental procedures for precise measurements of volumetric expansion and phase equilibria are provided elsewhere [41, 42]. As such, liquids are divided into three general classes [43]. Class I liquids such as water have insufficient ability to dissolve CO_2 , and therefore do not expand significantly. Glycerol and other polyols also fall into this class (Fig. 1a).

Class II liquids, such as ethyl acetate, acetonitrile, methanol, and hexane, dissolve large amounts of CO_2 and hence expand appreciably (Fig. 1) undergoing significant changes in physical properties. As shown in Fig. 1, conventional solvents such as acetonitrile and ethyl acetate are volumetrically expanded several fold by compressed CO_2 at 40°C , reaching as high as eight for ethyl acetate at around 70 bar. Regardless of the solvent, the volumetric expansion of class II solvents is strongly dependent on the mole fraction of CO_2



Gas Expanded Liquids for Sustainable Catalysis. Figure 1

(a) Expansion of solvents as a function of the pressure of CO_2 at 40°C , for ethyl acetate (●), MeCN (Δ), [1-butyl-3-methylimidazolium] BF_4 (filled squares), polypropylene glycol (□), and polyethylene glycol (○) [1]. (b) Vapor-liquid equilibrium for $\text{C}_3\text{H}_6 + \text{MeOH}$ binary system at 21°C [46]

in the liquid phase [41]. Class III liquids, such as ionic liquids and liquid polymers, dissolve relatively smaller amounts of CO₂ and therefore expand only moderately (Fig. 1a, [44]). By using a solvent such as acetonitrile or methanol that exhibits mutual solubility in CO₂ and water, it is possible to create CO₂-expanded ternary systems containing water [45]. Gases such as propylene can also expand organic solvents. Figure 1b shows that for the propylene + methanol binary system, the propylene mole fraction in the liquid phase approaches approximately 30 mol% at 10 bar, implying that compressed propylene at ambient temperatures can cause significant volumetric expansion of the methanol phase. Such behavior has been exploited in epoxidation of light olefins, as explained in a later section.

The expansions of Class II liquids have been successfully modeled by the Peng-Robinson Equation of State (PR-EoS) [47] and by molecular simulations [48, 49]. Gases such as ethane, fluoroform, and other similar compressible gases are also capable of expanding liquids. As discussed in later sections of this entry, gaseous substrates (such as propylene and ethylene) and gaseous oxidants (ozone) have also been exploited as expansion gases to overcome solubility limitations in the liquid phase where reaction occurs [50].

The presence of CO₂ in the liquid phase affects the physicochemical properties. For example, the viscosity of methanol decreases with CO₂ pressure at 40°C, by nearly 80% from pure methanol to CXL methanol at 77 bar CO₂ [51]. This viscosity reduction is especially striking for ionic liquids, which often have higher viscosity than organic solvents [52, 53]. The diffusivity of benzene in CXL methanol at 40°C and 150 bar increases by over 200% on replacing pure methanol with 75% CO₂ [54]. Similarly, the diffusion coefficients of benzonitrile in CO₂-expanded ethanol were observed to increase with increasing CO₂ mole fraction [55].

As regards polarity, Roškar et al. [56] have measured the dielectric constant of methanol with CO₂ at 35°C and found that from ambient conditions to approximately 40 bar of CO₂, the dielectric constant changes very little. However, the dielectric constant significantly decreases by about an order of magnitude at 76 bar of CO₂ pressure. It has been shown that the Kamlet–Taft parameters (for acidity, basicity, and polarizability) for mixed CO₂ + organic solvents are tunable by CO₂ addition [57–59]. This tunability has

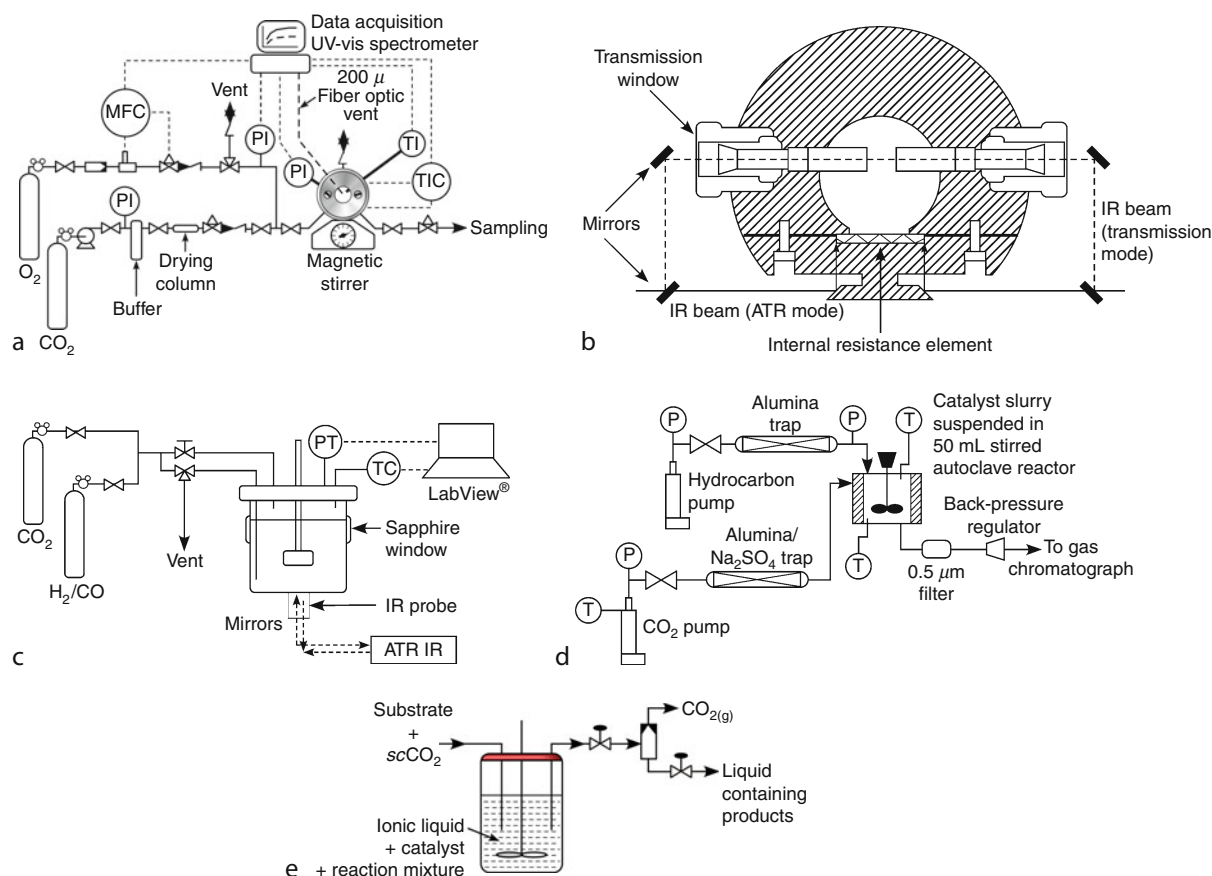
potential ramifications for the stabilization of charged and polar compounds in CXLs. The following sections provide examples of how GXLs have been systematically exploited to develop greener process concepts that display improved reaction performance with respect to rates, selectivity, and separations.

Catalytic Reactors for Investigating GXLs

Either batch or continuous-flow reactors may be employed for performing catalytic conversion, selectivity, and kinetic studies. In all cases, it is essential to know the phase behavior of the reaction mixture to rationally interpret the results. Equipment must be designed and pressure-tested according to standardized design and testing procedures such as those prescribed by the American Society for Testing Materials. All equipment must incorporate adequate pressure relief and inherent safety measures. For example, in the case of exothermic reactions, the amount of reactants fed should be such that the estimated adiabatic temperature rise at total conversion does not lead to “thermal runaway” conditions and unsafe pressures. Similarly, enough inerts should be added to the vapor phase to avoid the formation of explosive organic vapors in the presence of air. During continuous reactor operation, safety shutdown measures should be incorporated. Described below are examples of batch and continuous reactors reported in the literature.

Batch Reactors

Proof-of-concept batch studies are typically done in 5–10 mL view cells, as shown in Fig. 2, typically rated to operate at 150°C and 200 bar. A schematic drawing of an experimental reactor is shown in Fig. 2. The reactor is a low-volume (10 mL) hollow cylinder with sapphire windows at each end, sealed by o-rings and screw caps. The sapphire windows allow visual inspection of the cell contents and permit in situ spectroscopic studies. The cell body has as many as five ports. Two of the ports are used for introducing reactants such as O₂ or H₂ and the compressed gas medium such as CO₂. Oxygen or hydrogen is typically introduced via a mass flow controller. The third port is used for injecting liquid reactant into the reactor and is connected to a safety head containing a rupture disk. A pressure transducer that continuously monitors the



Gas Expanded Liquids for Sustainable Catalysis. Figure 2

Schematics of experimental reactors used for investigations of catalytic reactions in CXLs: (a) 5–10 mL view cell; (b) ATR-IR view cell; (c) 50 mL stirred reactor with ReactIR (PT pressure transducer, TC thermocouple); (d) continuous CXL reactor; (e) continuous flow reactor using a supercritical fluid–ionic liquid biphasic system

reactor pressure and a thermocouple that monitors the reactor temperature are connected to the remaining two ports. A magnetic stirrer provides adequate mixing of the reactor contents. Fiber optics may be attached to the sapphire windows and connected to a UV-Vis spectrophotometer. These facilitate temporal in situ monitoring of chemical reactions over broad spectral ranges. At the end of a batch run, the fluid phase cell contents may be depressurized into suitable sample traps for off-line analysis.

Baiker and coworkers [60] employed in situ high-pressure attenuated total reflectance infrared (ATR-IR) spectroscopy to elucidate the molecular interactions between dissolved CO_2 and ionic liquids. This cell, shown in Fig. 2c, consists of a horizontal stainless

steel cylinder with sapphire windows on each end, one of which is attached to a piston to control the reactor volume and pressure [61]. It can operate at temperatures up to 100°C and pressures up to 20 bar. The IR beam can be directed by a combination of four mirrors either through the ZnSe internal resistance element for ATR-IR measurements of the dense phase or through cylindrical ZnSe windows for transmission spectroscopy of the upper phase. In addition, the internal resistance element can be coated with a layer of catalyst to measure the interactions between the reactants and the catalyst during the reaction. Thus, the cell can be used to simultaneously provide information about dissolved and adsorbed reacting species and by-products as well as the catalyst itself.

For larger-scale batch investigations, adequate mechanical mixing must be ensured to avoid mass transfer limitations. Mechanically stirred reactors (50–300 mL such as Autoclave and Parr) with stirrer speeds up to 1500 rpm are well suited for this purpose. For example, a 50-cm³ high-pressure autoclave reactor equipped with an in situ attenuated total reflectance (ATR) IR probe (Mettler Toledo Inc.) was employed to investigate the hydroformylation of 1-octene [62]. Figure 2c shows a schematic of the apparatus (entitled “ReactIR”). Mixing is provided by a magnetic stirrer with a maximum agitation rate of 1700 rpm. Pressure and temperature are monitored by a Labview® data acquisition system and controlled with a Parr 4840 controller. Syngas is introduced from a gas reservoir, which is equipped with a pressure regulator that is used to admit syngas and maintain a constant total pressure in the reactor. The pressure transducer monitors the total pressure in the reservoir. The IR probe, placed at the bottom of the reactor, monitors concentration profiles of various species in the CXL phase. The maximum working pressure of the probe is approximately 140 bar. The temporal IR data are then processed using ConcIRT software to extract the absorbance profiles for each species, based on their characteristic peaks (identified with standards).

Continuous Reactors

For performing heterogeneous catalytic reactions in GXLs, a continuous stirred tank reactor (CSTR) is better suited for isothermal pressure-tuning studies [63]. A schematic is shown in Fig. 2d wherein the experiments are conducted in a 50 mL reactor from Autoclave Engineers, rated to 344 bar and 350°C. Catalyst particles are suspended in the reaction mixture by an impeller operating at 1200 rpm. Reaction pressure is maintained with a dome-loaded back-pressure regulator.

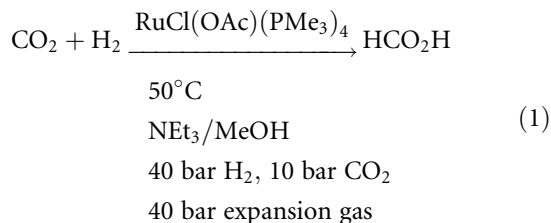
Cole-Hamilton and coworkers [64] demonstrated a continuous reactor system for investigating homogeneous catalytic reactions in IL/*sc*CO₂ systems (Fig. 2e). Here, the IL phase containing dissolved catalyst is retained in the reactor. The *sc*CO₂ is used to transport the soluble substrate into and products from the IL phase continuously. The

challenge is to minimize catalyst leaching from the IL phase by the *sc*CO₂ + reaction mixture exiting the reactor.

Applications in Homogeneous Catalysis

Hydrogenations

Hydrogenations have been shown to benefit from CO₂ expansion of the solvent in several ways. For example, CO₂ + H₂ mixtures are just as effective chemically as pure H₂ at the same total pressure, and also desirable from a safety standpoint. Jessop’s group [65] showed that the properties of the expanding gas can have a major effect on reaction rate. Investigating the effect of expansion gas on the rate of CO₂ hydrogenation in MeOH/NEt₃ mixture as solvent (Eq. 1), it was found that the turnover frequency was 770 h^{−1} with no expansion gas but 160 h^{−1} with ethane (40 bar) and 910 h^{−1} with CHF₃ (40 bar) as the expansion gas. The addition of liquid hexane produced a similar decrease in the rate. Hence, the low polarity of ethane was implicated for the decreased rate in the ethane-expanded solvent.

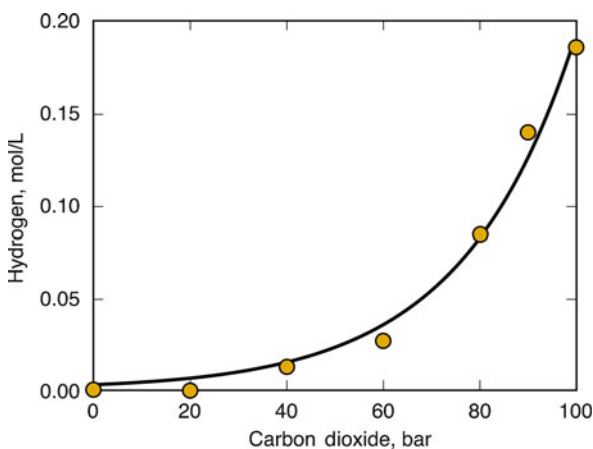


Solvent expansion as a method to enhance the H₂ availability in the liquid phase of homogeneous catalytic reactions is of interest in asymmetric hydrogenations. Foster’s group [66] investigated the asymmetric hydrogenation of 2-(6′-methoxy-2′-naphthyl)acrylic acid, an atropic acid, in CO₂-expanded methanol with [RuCl₂(BINAP)(cymene)]₂, finding the reaction faster but less selective than in neat methanol. A subsequent study using RuCl₂(BINAP) catalyst reported that the reaction in expanded methanol was slower than in normal methanol [67]. More recently, the asymmetric hydrogenation of methyl acetoacetate to methyl (R)-3-hydroxybutyrate by [(R)-RuCl(binap)(*p*-cymen)]Cl was investigated in methanol-dense CO₂ solvent systems [68]. Although the CO₂-expanded methanol system resulted in a reduction of both reaction rate and

product selectivity, this changed in the presence of water. High selectivities were obtained with the optimized methanol–CO₂–water–halide system.

Contrasting effects are also demonstrated during asymmetric hydrogenations in ionic liquids (ILs). The enantioselectivity during hydrogenation of tiglic acid in [BMIm][PF₆] (where BMIm is 1-*n*-butyl-3-methylimidazolium) is superior [H_2 = 5 bar; *ee* = 93%] to that in CO₂-expanded [BMIm][PF₆] (H_2 = 5 bar; CO₂ = 70 bar; *ee* = 79%) wherein the H_2 availability is improved with CO₂ [69]. In contrast, the hydrogenation of atropic acid is greatly improved in selectivity (*ee* increases from 32% to 57%) when the IL is expanded with 50 bar CO₂ [70]. Leitner and coworkers [69] demonstrated that hydrogenation of *N*-(1-phenylethylidene)aniline proceeded to only 3% conversion in [EMIm][Tf₂N] (1-ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide) and to >99% in CO₂-expanded [EMIm][Tf₂N], in which the H_2 solubility is significantly enhanced with CO₂ pressure (Fig. 3).

Facile hydrogenation of solid substrates such as vinyl naphthalene with a RhCl(PPh₃)₃ catalyst was demonstrated by melting the solid with compressed CO₂ and performing the reaction in the melt phase at 33°C, which is well below the normal melting point of the solid [71]. Scurto and Leitner [72] showed that the



Gas Expanded Liquids for Sustainable Catalysis.

Figure 3

Enhancement of H_2 solubility in [EMIm][Tf₂N] by CO₂ pressurization at a constant H_2 partial pressure (30 bar at T = 297 K). Taken from [69]

addition of compressed CO₂ induced a melting point depression of an ionic solid salt (tetrabutylammonium tetrafluoroborate) at greater than 100°C, well below its normal melting point of 156°C. They used this CO₂-induced melting technique to demonstrate hydrogenation, hydroformylation, and hydroboration of vinyl naphthalene using rhodium complexes. Employing homogeneous octene hydrogenation with rhodium complexes as a model system, Scurto and coworkers demonstrated the importance of understanding the phase behavior, mass transfer, and intrinsic kinetics effects for the reliable interpretation and design of hydrogenations in IL + CO₂ media [73].

Hydroformylations

Hydroformylations benefit from CXLs by virtue of the fact that syngas + CO₂ mixtures are more effective than syngas alone at a fixed pressure. The use of CO₂ helps generate CXLs, which enhance both the syngas solubility as well the tunability of the H_2 /CO ratio at milder pressures. Jin and Subramaniam [74] demonstrated the advantages of homogeneous hydroformylation of 1-octene in CXLs using an unmodified rhodium catalyst (Rh(acac)(CO)₂). The turnover numbers (TONs) in CO₂-expanded acetone were significantly higher than those obtained in either neat acetone or *sc*CO₂. In subsequent reports, the performances of several rhodium catalysts, Rh(acac)(CO)₂, Rh(acac)[P(OPh)₃]₂, Rh(acac)(CO)[P(OAr)₃], and two phosphorous ligands, PPh₃ and biphosphos, were compared in neat organic solvents and in CXLs [75]. For all catalysts, enhanced turnover frequencies (TOFs) were observed in CXLs. For the most active catalyst, Rh(acac)(CO)₂ modified by biphosphos ligand, the selectivity to aldehyde products was improved from approximately 70% in neat solvent to nearly 95% in CXL media. Such improved catalyst performance for hydroformylation in CO₂-based reaction mixtures were also reported recently for rhodium catalysts modified with triphenylphosphine, triphenyl phosphite, and tris(2,4-di-*tert*-butylphenyl) phosphite [76], with turnover numbers as high as 3(10⁴) mol aldehyde/(mol Rh)/h. In contrast, the Rh complex-catalyzed hydroformylation of 1-hexene in CO₂-expanded toluene was more rapid than in *sc*CO₂ but slower than in normal toluene [77]. The high CO₂ pressure was believed to

shift some of the 1-hexene out of the liquid phase into the CO₂ phase, thereby lowering the concentration of hexene available to the catalyst.

Ionic liquid + CO₂ media has also been used for hydroformylation [64, 78]. Constant activity for up to 3 days was demonstrated during continuous 1-octene hydroformylation with a Rh-based catalyst in [BMIm][PF₆]/scCO₂ medium. Compared to the industrial cobalt-catalyzed processes, higher TOFs were observed; the selectivity to linear aldehyde (70%) is comparable to those attained in the industrial processes (70–80%). However, the air/moisture sensitivity of the ILs and the ligands may lead to the deactivation and leaching of the rhodium catalyst. Cole-Hamilton and coworkers also reported a solventless homogeneous hydroformylation process using dense CO₂ to transport the reactants into and transfer the products out of the reactor leaving behind the insoluble Rh-based catalyst complex in the reactor solution [79].

Table 1 compares the TOFs, *n/i* ratio, and operating conditions (P&T) of some existing industrial processes and recently published work that employ CXLs, supercritical CO₂, and ionic liquids as reaction media [64]. Clearly, hydroformylation in CO₂-expanded octene appears to be promising in terms of both TOF (~300 h⁻¹) and selectivity (~90%, *n/i* > 10). In addition, the required operating conditions (60°C and 38 bar) are much milder compared to other processes. The further development of this process, including quantitative sustainability analysis, is provided in a following section.

Another example involves enantioselective hydroformylation of solid vinyl naphthalene with a Rh catalyst in a CO₂-based phase, created by melting the solid with

compressed CO₂. This technique is similar to the melt hydrogenation described earlier.

O₂-Based Oxidations

CXLs provide both rate and safety advantages for homogeneous catalytic oxidations. For the homogeneous catalytic O₂-oxidation of 2,6-di-*tert*-butylphenol, DTBP, by Co(salen*) in scCO₂, in CO₂-expanded acetonitrile and in the neat organic solvent [80, 81], it was found that the TOF in the CO₂-expanded acetonitrile is between one and two orders of magnitude greater than in scCO₂. Additionally, the 2,6-di-*tert*-butyl quinine (DTBQ) selectivity is lower in neat acetonitrile, clearly demonstrating that CXLs are optimal media for this reaction. In another example, cyclohexene oxidation by O₂ was investigated with a non-fluorinated iron porphyrin catalyst, (5,10,15,20-tetraphenyl-21*H*,23*H*-porphyrinato)iron(III) chloride, Fe(TPP)Cl, and a fluorinated catalyst (5,10,15,20-tetrakis(pentafluorophenyl)-21*H*,23*H*-porphyrinato)iron(III) chloride, Fe(PFTPP)Cl [80]. While Fe(TPP)Cl is insoluble and displays little activity in scCO₂, it displays high activity and selectivity in CO₂-expanded acetonitrile. The enhanced TOFs in the CXL media were attributed to their tunable polarity compared to scCO₂ and the approximately two orders of magnitude greater O₂ solubility in CO₂-expanded solvents compared to neat solvents at ambient conditions (1 atm and 25°C). These results clearly show that CO₂-expanded solvents advantageously complement scCO₂ as reaction media by broadening the range of conventional catalyst/solvent combinations with which homogeneous oxidations by O₂ can be performed.

Gas Expanded Liquids for Sustainable Catalysis. Table 1 Comparison with commercial processes and other reported work

Process parameters	BASF (Co)	Shell (Co/P)	SCF-IL (Rh/P)	SCF (Rh/P)	CXL (Rh/P)
Substrate	1-octene	1-octene	1-octene	1-octene	1-octene
P, bar	300	80	200	125	38
T, °C	150	200	100	100	60
TOF, h ⁻¹	35	20	517	259	316
S _n -aldehyde, %	50	80	75	75	89

An important safety aspect when using CXLs is that the dominance of dense CO₂ in the vapor phase reduces the flammability envelope. Brennecke and coworkers showed that either pure O₂ or O₂ + CO₂ mixtures at the same total pressure showed more or less similar O₂ solubilities in pressurized acetonitrile or methanol [2]. Table 2 shows the vapor–liquid equilibrium data for the CO₂/O₂/acetonitrile system at approximately 40°C and pressures between 6 and 83 bar. The CO₂ and O₂ mole fractions are shown in the table, with the balance being acetonitrile. To better understand the effect of CO₂ presence on O₂ solubility in acetonitrile, the authors use the Enhancement Factor (EF), defined as the ratio of the solubility of O₂ in the liquid phase of the ternary mixture to the solubility of pure O₂ in the solvent at the same O₂ fugacity and temperature.

$$EF = \frac{x_{\text{gas}}^{\text{mixture}}}{x_{\text{gas}}^{\text{solvent}}}$$

Clearly, EF is >1 implies that the presence of the CO₂ increases the solubility of the O₂. As can be inferred from Table 2, the EF values increase with an increase in pressure and reached values greater than 1 at pressures above 50 bar. EF values >1 can be attributed to the high CO₂ solubility in the liquid phase, which

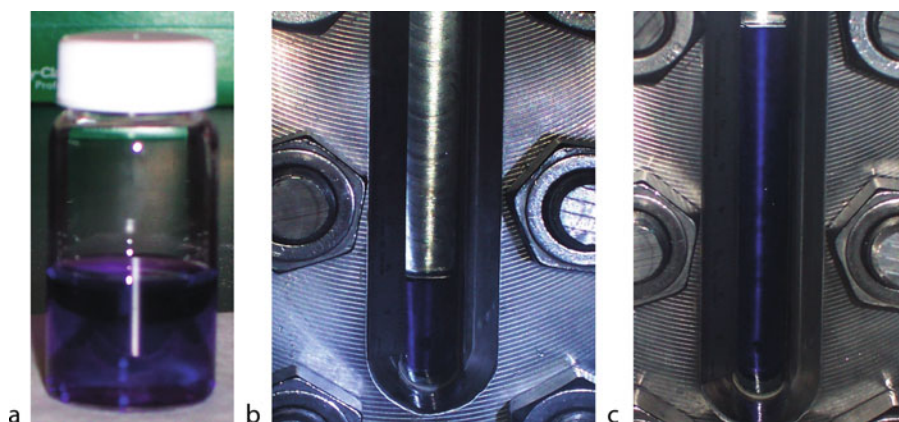
increases free volume in the solution and enhances the solubility of the O₂. If the binary and ternary systems are compared at the same total pressure (rather than the same O₂ fugacity), then the oxygen solubility in the ternary mixture liquid phase is lower than the solubility in the binary system. This finding is significant because it confirms that replacing CO₂ with O₂ in the gas phase does not drastically change the O₂ solubility in the CXL phase but serves to significantly mitigate vapor-phase flammability and thereby to enhance process safety.

The Mid-Century (MC) process for the oxidation of methylbenzenes to carboxylic acids represents an important industrial process for the synthesis of polymer intermediates for producing fibers, resins, and films. For example, terephthalic acid (TPA) is manufactured by homogeneous liquid-phase oxidation of *p*-xylene around 200°C and 20 bar with Co/Mn/Br catalysts in acetic acid medium. The air (source of oxygen) is vigorously sparged into the liquid phase of a stirred tank reactor. The purity of the solid TPA product obtained from this reactor is typically 99.5% pure, the major impurity being 4-carboxybenzaldehyde. For obtaining polymer-grade TPA, further purification is necessary to eliminate the intermediate oxidation products suggesting that the vigorous mixing does not completely overcome the O₂ availability limitations. Recently, the Co/Mn/Br catalyzed oxidation of *p*-xylene

Gas Expanded Liquids for Sustainable Catalysis. Table 2 Vapor–liquid equilibrium for CO₂ (1)–O₂ (2)–acetonitrile (3) [2]

<i>T</i> (°C)	<i>P</i> _{total} (bar)	Liquid-phase composition		Vapor-phase composition		<i>f</i> _{O₂} (bar)	<i>x</i> _{O₂} , pure gas ^a mole fraction	EF
		<i>x</i> ₁	<i>x</i> ₂	<i>y</i> ₁	<i>y</i> ₂			
40.0	13.7	0.03	0.004	0.27	0.68	9.2	0.0046	0.77
40.0	23.7	0.08	0.007	0.35	0.63	14.9	0.0074	0.92
40.0	26.8	0.08	0.009	0.30	0.68	17.9	0.0089	0.98
40.0	48.9	0.15	0.016	0.36	0.62	29.8	0.0148	1.05
40.0	69.4	0.20	0.022	0.37	0.61	41.6	0.0207	1.08
40.0	82.2	0.24	0.028	0.38	0.61	48.6	0.0242	1.18
40.0	6.2	0.02	0.001	0.40	0.52	3.2	0.0016	0.59
40.0	16.7	0.09	0.003	0.55	0.42	7.0	0.0035	0.86
39.8	29.6	0.16	0.005	0.60	0.38	11.1	0.0055	0.99
40.0	52.7	0.29	0.012	0.64	0.34	18.0	0.0089	1.31

^aHorstmann et al. 2004 [82]



Gas Expanded Liquids for Sustainable Catalysis. Figure 4

Example of CO_2 expansion of a homogenous oxidation catalyst mixture in acetic acid. (a) room temperature, (b) 120°C with 20 bar N_2 and 18 bar CO_2 , (c) 120°C with 20 bar N_2 and 159 bar CO_2 ; $[\text{Co}] = 60 \text{ mM}$, $[\text{Mn}] = 1.8 \text{ mM}$, $[\text{Br}] = 60 \text{ mM}$ prior to the expansion

to TPA in CO_2 -expanded solvents was demonstrated at temperatures lower than those of the traditional MC process [82]. As inferred from Fig. 4, the acetic acid solution containing the Co/Mn/Br-based catalyst undergoes significant volumetric expansion by CO_2 addition.

As compared with the traditional air (N_2/O_2) oxidation system, the reaction with CO_2/O_2 at 160°C at a pressure of 100 bar increases both the yield of TPA and the purity of solid TPA via a more efficient conversion of the intermediates, 4-carboxybenzaldehyde and *p*-toluic acid. Further, the amount of yellow colored by-products in the solid TPA product is also lessened. Additionally, the burning of the solvent, acetic acid, monitored in terms of the yield of the gaseous products, CO and CO_2 , is reduced by approximately 20% based on labeled CO_2 experiments. These findings show that the use of CXLs promotes sustainability by maximizing the utilization of feedstock carbon for desired products while simultaneously reducing carbon emissions.

H_2O_2 -Based Oxidations

Many of the current commercial propylene oxide (PO) processes are energy intensive and yield coproducts [83]. They either consume vast amounts of chlorine and lime producing large quantities of wastewater containing HCl and salt, or use expensive reactants that

lead to equimolar amounts of coproducts. A novel propylene epoxidation process has been disclosed wherein titanium-substituted silicalite (TS-1) catalysts [84–87] catalyze propylene epoxidation with reasonable efficiency using an O_2/H_2 mixture to generate H_2O_2 in situ. TS-1 catalysts have high catalytic activity and selectivity. However, the catalyst deactivates rapidly and requires high temperatures for regeneration.

Eckert/Liotta and coworkers showed that the reaction between H_2O_2 and dense CO_2 (benign reactants) yields a peroxycarbonic acid species, an oxidant that facilitates olefin epoxidation [88]. Using such a system with NaOH as a base, Beckman and coworkers [89] demonstrated propylene conversion to PO at high selectivity albeit at low conversion (3%). The low conversion is typical of interphase mass transfer limitations between the immiscible aqueous and CO_2 phases. By using a third solvent such as acetonitrile that shows mutual solubility with water and dense CO_2 , CO_2 -expanded $\text{CH}_3\text{CN}/\text{H}_2\text{O}_2/\text{H}_2\text{O}$ homogeneous mixtures were created. Subramaniam and coworkers [45] showed that olefin epoxidation reactions may be intensified in such a homogeneous phase containing the olefin, CO_2 , and H_2O_2 (in aqueous solution). One to two orders of magnitude enhancement in epoxidation rates (compared to the biphasic system without acetonitrile) was achieved with >85% epoxidation selectivity. This is an example of a system where the CO_2 is used as a solvent (for the olefinic

substrate) and as a reactant to generate the peroxy carbonic acid in situ. Recently, a similar concept was proposed for styrene epoxidation in CO₂-based emulsions [90].

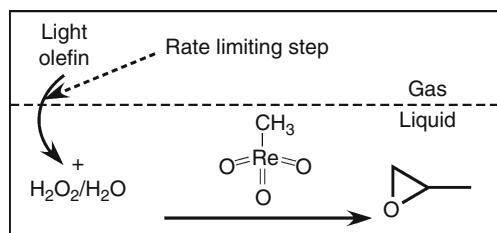
Propylene Epoxidation in Propylene-Expanded Liquid Phase Methyltrioxorhenium (CH₃ReO₃, abbreviated as MTO) is known to be an exceptional homogeneous catalyst for alkene epoxidation at relatively mild temperatures using H₂O₂ as the oxidant [91–94]. This catalytic system was recently shown to work elegantly for the industrially significant epoxidations of propylene [95] and ethylene [96]. As shown in Scheme 1, the light olefin (either propylene or ethylene) from the gas phase has to first dissolve into the aqueous liquid phase before undergoing reaction. However, the aqueous solubility of the olefins is typically low and limits the epoxidation rate. For example, propylene solubility in water is 1.36 (10^{−4}) M at 1 bar, 21°C [97]. The key to enhancing the reaction rate was to exploit the compressibility of the light olefins at reaction temperatures to enhance their solubilities in the liquid phase. By dissolving a suitable amount of methanol to the liquid phase, the solubility of the light olefins can be substantially enhanced.

For propylene epoxidation, it was found that if a N₂/C₃H₆ mixture at 14 bar was used in the gas phase, remarkably high activity (92% propylene oxide yield in 1 h) was achieved [95]. When pressurizing with N₂ in a closed system at ambient temperatures, C₃H₆ (P_c = 46.1 bar; T_c = 92.5°C) condenses at 14 bar [98]. In the presence of a solvent such as methanol,

pressurization beyond 10 bar increases the C₃H₆ mole fraction in the liquid phase to >30 mol% [46]. This increased concentration of propylene results in increased epoxidation rates. This is an example of how to utilize a light olefin as the expansion medium and exploit a *substrate-expanded liquid phase* for process intensification.

Table 3 compares the various GXL-based processes with the commercial chlorohydrin and hydroperoxide processes. The pressure-intensified propylene epoxidation process (last entry in Table 3) satisfies the sustainability principles of waste minimization, use of benign reagents, and process intensification at mild conditions. Little or no waste is produced compared to the commercial processes.

Ethylene Epoxidation in Ethylene-Expanded Liquid Phase Conventional ethylene oxide manufacture emits CO₂ as by-product (roughly 18 MM tons/year) from the combustion of both the ethylene (feed) and ethylene oxide (EO) [100]. An alternate technology that is exclusively selective toward EO (eliminating the formation of CO₂) would dramatically reduce the carbon footprint of this large-scale industrial process. It was recently shown that the MTO-based liquid-phase epoxidation process works remarkably well for ethylene as well. At ambient temperatures, the gaseous ethylene (P_c = 50.6 bar; T_c = 9.5°C) is just above the critical temperature. Hence, by compressing the ethylene gas beyond the critical pressure (>50 bar), it is possible to significantly increase its solubility (by one to two orders of magnitude) in a liquid reaction phase containing methanol or another suitable alcohol [101]. By employing MTO, H₂O₂, and a methanol/water mixture as solvent, a homogeneous catalytic system was demonstrated that eliminates CO₂ formation while producing ethylene oxide at >99% EO selectivity at near-ambient temperatures and EO productivities [~4 g EO/(g Re)/(h)] that are comparable to the conventional EO process (Table 4). No CO₂ was detectable in either the liquid or the vapor phases. Furthermore, since H₂O₂ does not decompose at typical reaction temperatures (<40°C), the vapor phase is void of O₂ and the formation of explosive vapors is impeded. In addition to mitigating the carbon footprint of a large-scale industrial process, the demonstrated technology concept is another example of how the



Gas Expanded Liquids for Sustainable Catalysis.

Scheme 1

Homogeneous catalytic epoxidation of propylene. The methyltrioxorhenium (MTO) catalyst and oxidant (H₂O₂) are dissolved in the liquid phase and the olefin is supplied from the gas phase

Gas Expanded Liquids for Sustainable Catalysis. Table 3 Comparison with other processes (substrate is propylene in all cases)

Process/reference	Reagents	Conditions	PO yield (Y) or selectivity (S)
Hancu et al. [89]	CO ₂ + H ₂ O ₂ + H ₂ O (Biphasic system), NaOH	–	Y < 3.0%
Danciu et al. [85]	CO ₂ + H ₂ + O ₂ , Pd/TS-1 catalyst	45°C; 131 bar	Y ~ 7.3%; S = 77%; 4.5 h
Biphasic CXL process, Lee et al. [99]	CO ₂ + CH ₃ CN + H ₂ O ₂ + H ₂ O (monophasic), pyridine	40°C; 48 bar	Y = 7.1%; 6 h
Chlorohydrin process, Trent [83]	Cl ₂ , caustic, lime	45–90°C; 1.1–1.9 bar	S = 88–95%
Hydroperoxide ^a process, Trent [83]	W, V, Mo	100–130°C; 15–35.2 bar	Y = 95%; S = 98%; 2 h
Pressure-intensified CXL process [95]	CH ₃ OH–H ₂ O ₂ –H ₂ O, MTO, pyNO	30°C; 17 bar N ₂	Y = +98%; S = 99%; 2 h

^aIsobutane to TBHP: 95 ~ 150°C; 20.7–55.2 bar; Y = 20–30%; S = 60–80%

Gas Expanded Liquids for Sustainable Catalysis. Table 4 Comparison of commercial EO process with new GXL-based homogeneous catalytic process

Process	Catalyst, oxidant	Pressure, temperature	EO selectivity	CO ₂ selectivity	EO productivity (g/gcat/h)
Shell	Ag/Al ₂ O ₃ , air or O ₂	10–20 bar, 200–300°C	85–90%	10–15%	3–6
GXL based	MTO, H ₂ O ₂	10–50 bar, 20–40°C	99+%	No CO ₂ detected	~4.5

synergy afforded by the facile compressibility of a substrate such as ethylene and the accompanying enhanced solubility in low molecular weight alcohols can be exploited to enhance selectivity and productivity.

Acid Catalysis

The replacement of mineral acids with benign and less hazardous alternatives has long been a grand challenge in chemicals synthesis. In situ generation of acids by the reaction of CO₂ with alcohols or water is desirable because depressurization leads to self-neutralization [102–104]. The formation of the dimethyl acetal of cyclohexanone is up to 130 times faster in CO₂-expanded methanol than in normal methanol without any added acid. Such in situ acids also catalyze the hydrolysis of β-pinene to terpineol and other alcohols with good selectivity for alcohols rather than hydrocarbons [103].

Addition of CO₂ to pressurized hot water accelerates reactions that can proceed by acid catalysis. CO₂ dissolved in water at 250°C promotes the decarboxylation of benzoic acid [105], the dehydration of cyclohexanol to cyclohexene, and the alkylation of p-cresol to 2-*tert*-butyl-4-methylphenol [106]. The hydration of cyclohexene to cyclohexanol at 300°C showed a fivefold rate increase as the CO₂ pressure was increased from 0 to 55 bar.

Miscellaneous

Ozonolysis Ozone has high oxidation potential ($E^\circ = 2.075$ V in acid and 1.246 V in base) and has been extensively investigated as a potent oxidant. Though considered toxic, ozone does not persist in the environment eventually decomposing to molecular oxygen. Ozone is effective for the cleavage of carbon–carbon double bonds. This oxidation reaction

is believed to proceed via metastable intermediates that upon further catalytic oxidation or reduction yield products that are suitable as building blocks for chemical synthesis. For example, the ozonolysis of unsaturated fatty acids can yield a range of both monoacids and diacids [107–110]. Ozone attacks most common organic solvents used as reaction media, creating undesirable waste products and consuming the ozone away from the desired reaction. It has been recently shown that ozone can be used effectively in liquid CO₂, in which it is not only stable but also remarkably soluble [111]. At typical ozonolysis temperatures (0–20°C), ozone is sufficiently close to its critical temperature (–12.1°C) such that the ozone density can be increased to liquid-like values by compression beyond its critical pressure (55.6 bar). Conveniently, in the 0–20°C range and beyond 50 bar, the CO₂ ($P_c = 73.8$ bar; $T_c = 31.1^\circ\text{C}$) exists as liquid and the O₃ content in this liquid phase is easily tuned with pressure. An order of magnitude increased dissolution of O₃ in liquid CO₂ creates an *O₃-expanded liquid phase*. The O₃ half-life in liquid CO₂ was found to be approximately 6 h at –1.2°C [111]. Further, substrates such as methyl oleate and *trans*-stilbene when dispersed in liquid CO₂ (containing dissolved O₃) undergo complete conversion in minutes to the corresponding aldehyde and acid products (nonanal, nonanoic acid in the case of methyl oleate; and benzaldehyde, benzoic acid in the case of *trans*-stilbene). The foregoing results clearly show that ozonolysis in liquid CO₂ is a facile, clean, and inherently safe oxidation route.

Carbonylations Exploiting the enhanced and tunable CO solubility in CXLs, it was recently shown that selective mono or double carbonylations could be achieved by using CO₂-expanded liquids during [2 + 2 + 1] carbonylative reactions of alkenes or acetylenes with allyl bromides catalyzed by Ni(I) [112].

Polymerizations Catalytic chain transfer polymerizations are free radical polymerizations that use a homogeneous catalyst to terminate one chain and initiate a new one. During the polymerization of methyl methacrylate, the chain transfer step is believed to be a diffusion-controlled reaction in which the Co(II) catalyst abstracts a hydrogen atom from the

polymer radical (R●) and transfers it to a monomer to start the growth of a new chain. Zwolak et al. [113] reported that the rate of chain transfer was fourfold greater in CO₂-expanded methyl methacrylate (60 bar, 50°C) than in neat monomer. The improvement was attributed to the lower viscosity of the CO₂-expanded solution.

Biomass Conversions Plant-based biomass has the potential to completely displace fossil fuel-based feedstocks for chemicals production in a sustainable manner. However, new catalytic technologies are needed for the fledgling biorefinery to convert plant-based biomass feedstocks to chemicals and chemical building blocks. Recently, the acid-catalyzed transesterification of soybean flakes in CO₂-expanded methanol containing sulfuric acid was demonstrated for producing fatty acid methyl esters (FAME) [114]. The authors report that the introduction of CO₂ into the system increases the rate of reaction by as much as 2.5 fold in comparison to control reactions without CO₂.

1,2-glycerol carbonate was formed from glycerol and carbon dioxide in methanol using (Bu₂SnO)-Bu-*n* (dibutyltin(IV)oxide, 1) as a catalyst [115]. The yield of 1,2-glycerol carbonate was as high as 35%. The reaction proceeds upon activation of the catalyst by methanol forming dibutyltin dimethoxide followed by dibutyltin glycerate, which undergoes CO₂ insertion to ultimately yield glycerol carbonate.

Multiphase Catalysis

The attractive properties of GXLs are also applicable in heterogeneous catalysis. For example, adding CO₂ to an organic liquid phase in a fluid–solid catalytic system should enhance gas solubilities and improve the mass transfer properties of the expanded liquid phase. Reviews of supercritical phase heterogeneous catalysis may be found elsewhere [116–118].

Hydrogenations

Employing dense CO₂ as the solvent medium in a stirred reactor, it was reported that the Pd/Al₂O₃ catalyzed hydrogenation of an unsaturated ketone proceeded faster in CO₂-expanded ketone than in the unswollen ketone [119]. A similar observation was reported during the Pd/C catalyzed hydrogenation of

pinene wherein the reaction rate was higher at lower pressures in a condensed CXL phase compared to single-phase operation at supercritical conditions [120]. Roberts and coworkers reported that the rate constant for the hydrogenation of the aromatic rings in polystyrene (PS) was found to be higher in CO₂-expanded decahydronaphthalene (DHN) than in neat DHN [121]. For the Pt/ γ -Al₂O₃ catalyzed hydrogenation of tetralin to decalin, Chan and Tan [122] reported enhanced rates in the presence of a CO₂-expanded toluene phase in a trickle bed reactor. For the Pd/C-catalyzed hydrogenation of CO₂-expanded α -methylstyrene, it was shown that the rate-enhancing effect of CO₂ is influenced by two competing factors: solvent strength and reactant concentration [123]. The presence of CO₂ modifies the solvent strength of the liquid phase, resulting in more favorable adsorption equilibrium for the surface reaction. However, the diluting effect of CO₂ leads to reduced reaction rates. Thus, there exists an optimum CO₂ level in the liquid phase that is tunable by reactor pressure.

During the NiCl₂-catalyzed reduction of benzonitrile to benzylamine by NaBH₄, Xie et al. [124] showed that CO₂ expansion of the reaction mixture converts the primary amine to a carbamate salt, thereby preventing its further reaction to secondary amines. The carbamic species release CO₂ upon gentle heating. Thus, the benzylamine yield was 98% in CO₂-expanded ethanol but <0.01% in normal ethanol.

Heterogeneous catalysis without solvent relies on both the substrate and product being liquids or gases. If the product is a solid at the reaction temperature, then the reaction will not proceed to completion because the reaction mixture will solidify before full conversion is obtained. Normally, this problem is solved by adding a solvent or using an elevated temperature, but a third option is to lower the melting point of the product by expansion with CO₂ [71]. For example, the Pt-catalyzed hydrogenation of oleic acid at 35°C stalls at 90% conversion even with extended reaction times (25 h). However, in the presence of 55 bar CO₂, the reaction proceeds to 97% conversion after only 1 h.

Selective Oxidations

During the Pd/Al₂O₃ catalyzed partial oxidation of octanol, Baiker and coworkers report significantly

enhanced oxidation rates at intermediate pressures where a condensed CXL phase probably exists compared to higher pressure where a single supercritical phase exists [125]. The reaction in the condensed phase benefits from increased concentrations of the substrate relative to the single supercritical phase while also enjoying adequate O₂ availability in the liquid phase.

During the O₂-based oxidation of cyclohexene on a MCM-41 encapsulated iron porphyrin chloride complex, conversion and selectivity in CO₂-expanded acetonitrile (~30 mol% CO₂) almost doubled compared to the neat organic solvent [126]. For the oxidation of 2,6-di-*tert*-butylphenol (DTBP) to 2,6-di-*tert*-butyl-1,4-benzoquinone (DTBQ) and 3,5,3',5'-tetra-*tert*-butyl-4,4'-diphenoquinone (TTBDQ), a series of porous materials with immobilized Co(II) complexes were screened as catalysts in neat acetonitrile, supercritical carbon dioxide (scCO₂), and CO₂-expanded acetonitrile [127]. In this case, the highest conversions were found in scCO₂ suggesting that scCO₂, rather than CXL or liquid reaction media, provides the best mass transfer of O₂ and of substrates through the porous catalysts.

Hydroformylation and Carbonylation

Abraham and coworkers [77] investigated 1-hexene hydroformylation over a rhodium–phosphine catalyst immobilized on a silica support and found that the rates in CO₂-expanded toluene and in scCO₂ were comparable but faster than in normal toluene. However, the activity declined with time due to possible catalyst leaching.

Leitner and coworkers have demonstrated that compressed CO₂ can be used to effectively overcome mass transfer limitations encountered during solid-phase organic synthesis with pressurized gaseous reagents [128]. Depending on the relative importance of mass-transfer limitations and catalyst/substrate concentration, CXLs may provide the optimum conditions for both hydroformylation and carbonylation reactions. For example, the catalytic carbonylation of norbornene (Pauson–Khand reaction) supported on a polymer support proceeds nearly quantitatively in CXL media. Here, the CXL medium provides the optimum combination of catalyst concentration and CO availability to maximize the reaction rate.

Solid Acid Catalysis

The acylation of anisole with acetic anhydride was investigated in a continuous slurry reactor over mesoporous-supported solid acid catalysts such as Nafion® (SAC-13) and heteropolyacids [129]. The CXL media gave lower conversion and, surprisingly, faster deactivation compared to a liquid-phase reaction despite the use of polar cosolvents. The deactivation is possibly due to retention of heavy molecules (possibly di- and tri-acylated products) formed by the interaction of acetic anhydride with para-methoxyacetophenone (*p*-MOAP) in the catalyst micropores. The addition of CF₃CO₂H has been shown to catalyze a Friedel–Crafts alkylation of anisole in CO₂-expanded anisole (95°C, 42 bar), but the reaction was no faster than that in normal anisole [130].

An Example of CXL-Based Process Development Including Sustainability (Economic and Environmental Impact) Analyses

For developing sustainable catalytic processes, a multiscale approach involving concurrent catalyst design, solvent engineering, and reactor engineering is essential [40]. An example of such an approach, involving chemists and engineers, is presented in this section for an industrially significant reaction.

For the hydroformylation of higher olefins, cobalt-based catalysts are employed. The cobalt catalysts require rather harsh operating conditions (140–200°C, 50–300 bar) and the catalyst recovery steps involve much solvents, acids, and bases [131]. The challenges for a sustainable technology alternative are to develop a process that operates at milder temperatures (<100°C) and pressures (<100 bar), and requires a simple yet environmentally friendly catalyst recovery method. The use of a Rh catalyst for 1-octene hydroformylation in CXL media provides exceptional TOF (~316 h⁻¹) and regioselectivity (*n/i* ~ 9) at very mild pressure (~40 bar) and temperature (30–60°C) compared to conventional Co-based processes [75]. This markedly enhanced regioselectivity in CXLs during homogeneous 1-octene hydroformylation is partly attributed to the beneficial tunability of the H₂/CO ratio in the CXL phase.

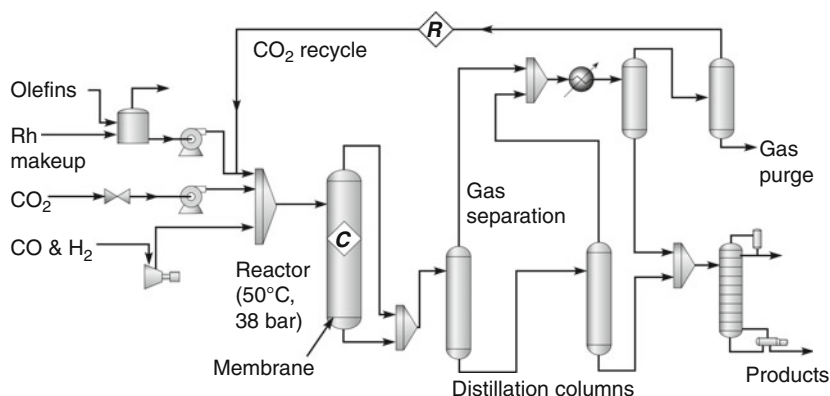
It is well known that the concentrations of the syngas components (CO and H₂) in the liquid phase

are major determinants of the reaction pathways and therefore the product selectivity. In general, higher H₂ concentrations are needed for catalyst activation and lower CO concentrations are required to achieve higher rates and avoid inhibition effect due to formation of inactive carbonyl species [132]. Because CO is generally more soluble than hydrogen in most conventional solvents [133], the resulting H₂/CO ratio in the liquid phase is less than that in the feed syngas. However, when CO₂ is added to either 1-octene or nonanal (to create a CXL), it was observed that the H₂ is more soluble than CO in the CXLs [5]. This means that the H₂/CO ratio in the liquid phase should be greater in CXLs (based on the organic solvent and extent of CO₂ addition) compared to the ratio in the feed.

Gas solubility measurements showed the presence of CO₂ at hydroformylation conditions (*T* = 40–80°C and pressures up to 90 bar) enhances the solubilities of both CO and H₂ in the liquid phase. The enhancement factor, defined as the ratio of the gas (CO or H₂) mole fraction in the neat solvent relative to that in the CXL at identical temperature and gas (CO or H₂) fugacities in the vapor phase, is greater for hydrogen (around 1.8) compared to carbon monoxide (around 1.5). This unique tunability of the H₂/CO ratio in CXL media is believed to enhance both the TOF and *n/i* ratio.

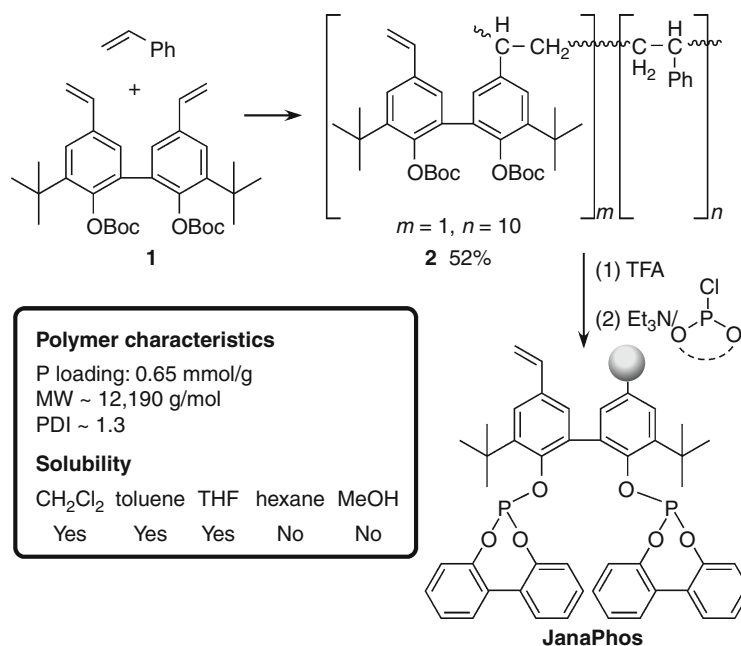
A detailed engineering model that takes into account kinetics, phase equilibrium, and mass transfer effects provided a better understanding of the mass transfer and kinetic effects in the CXL-based media [134]. A plant-scale simulation of the CXL process concept (Fig. 5) was constructed to facilitate economic and environmental impact analyses. Economic analysis revealed that >99.8% rhodium has to be recovered per pass for the CEBC hydroformylation process to be competitive with a simulated Co-based commercial process [135]. Environmental impact analysis revealed that the CEBC process produces half as much waste with lower overall toxicity compared to the simulated conventional process [135].

To develop Rh catalysts that meet the quantitative criterion for economic viability, a soluble polymer-attached, recyclable rhodium(I) catalyst with chelate-capable phosphite functionality, that was used to produce a polymer ligand, was synthesized (Scheme 2). By controlling the molecular weight, the polymer is designed such that it is completely soluble in



Gas Expanded Liquids for Sustainable Catalysis. Figure 5

Process flow diagram for CXL-based hydroformylation concept

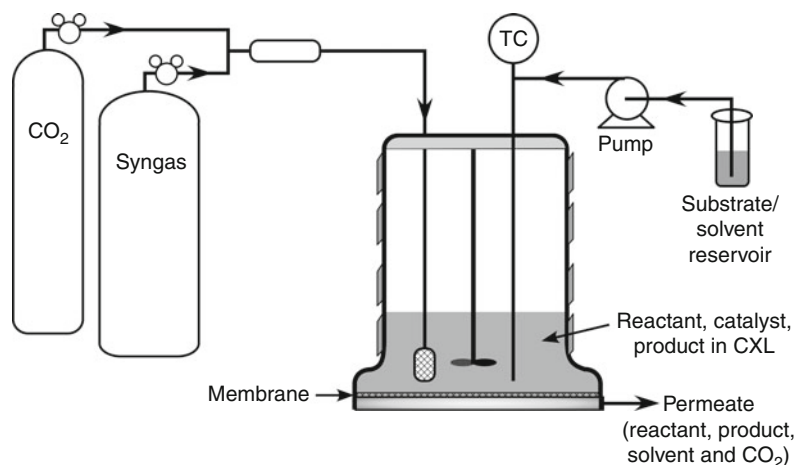


Gas Expanded Liquids for Sustainable Catalysis. Scheme 2

Synthesis of polymer-attached ligand [136]

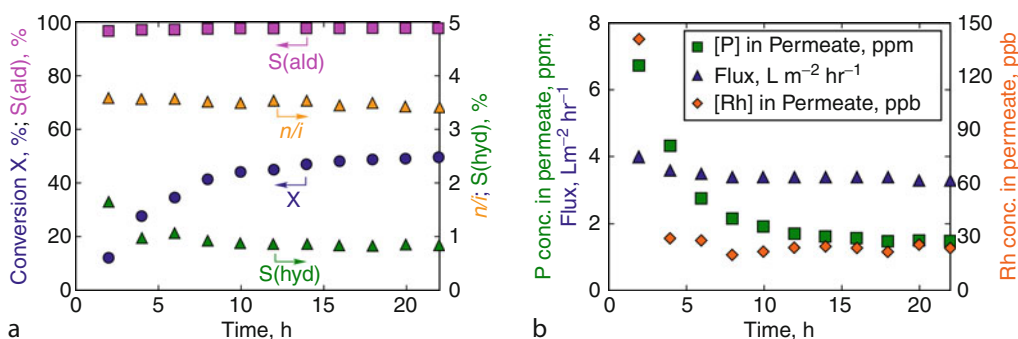
the hydroformylation reaction medium yet bulky enough to diffuse through a nanofiltration membrane [136, 137]. The polymer support was designed to bind Rh in a bidentate fashion to provide better site isolation for the rhodium catalysts as well as to inhibit decomplexation and subsequent leaching of rhodium from the polymer.

Using the soluble polymer-attached ligand to bind the Rh precursor, a continuous hydroformylation process concept that uses Rh-based homogeneous catalysts and operates at mild pressures (tens of bar) and temperatures less than 100°C has been demonstrated. The reactor schematic is shown in Figure 6. Syngas and compressed CO₂ (to generate CXLs) are



Gas Expanded Liquids for Sustainable Catalysis. Figure 6

Schematic of continuous homogeneous hydroformylation in CXL with catalyst retention by membrane nanofiltration



Gas Expanded Liquids for Sustainable Catalysis. Figure 7

Demonstration of a continuous CXL-based homogeneous hydroformylation reactor with effective catalyst retention by membrane nanofiltration. $T = 50^\circ\text{C}$, $P_{\text{syngas}} = 30$ bar, $\text{CO}/\text{H}_2 = 1:1$, $\text{P}/\text{Rh} = .6$, $\text{Oct}/\text{Rh} = 2200$, toluene/1-octene = 70:30 by volume, stirrer speed: 1000 rpm [138]

added to the 1-octene + nonanal reaction mixture in which the polymer-attached rhodium complex is dissolved. The product stream is continuously withdrawn while maintaining the reactor pressure constant.

Continuous operation is characterized by steady membrane flux, constant conversion, and constant product selectivity (Fig. 7a). At steady state, the permeate stream contains the unreacted 1-octene and CO_2 (which is separated by depressurization and may be recycled) products. As seen from Fig. 7b, the Rh and P concentrations in the permeate phase, quantified

using ICP analysis, were on the order of a few tens of ppb [138]. The cost of the makeup Rh is 0.4 cent/lb product, which exceeds the economic viability criterion.

The foregoing example represents a systems-based, multiscale research approach developing novel, environmentally beneficial, and economically viable process concepts. A patent pertaining to the CXL-based hydroformylation concept [139] has been licensed to a company for a defined field of use. The demonstrated technology concept, when fully optimized, should find applications in a variety of other applications in

homogeneous catalysis, including hydrogenation and carbonylation of conventional and biomass-based substrates.

Summary and Future Directions

The novel science and technology advances discussed in this entry demonstrate the various unique ways in which GXL media may be exploited to develop greener process concepts for oxidations, hydroformylations, hydrogenations, ozonolysis, and other chemistries. The demonstrated advantages include *process intensification* at mild conditions by increasing dissolution of the limiting reagent in the GXL reaction phase; the *efficient utilization* of feedstock and reactive gases such as O_3 due to the inertness of CO_2 , an often used expansion medium; enhancing *inherent safety* of the process by suppression of flammable vapors; and *waste minimization* by suppression of side reactions that generate undesired products such as CO_2 and reduced usage of volatile organic solvents.

The alternative concepts presented herein address chemistries underlying large-scale catalytic technologies, including the homogeneous hydroformylation of higher olefins and the epoxidations of light olefins such as ethylene and propylene. The global capacity of these processes is growing at 4–6% annually. Hence, the deployment of viable alternative technologies for future expansion of these processes (and also as replacement of existing units when needed) would have a significant impact in reducing environmental footprints. For these promising concepts to be developed further and considered for commercialization, continued fundamental investigations that integrate catalysis, phase behavior involving GXL media (for reactions and separations), kinetic and reactor modeling are essential. These investigations must be complemented by ongoing quantitative economic and environmental impact analyses to provide research and process engineering guidance for developing practically viable process concepts.

The use of renewable feedstocks has the potential to completely displace petroleum crude and coal for producing industrial chemicals. Further, reducing the carbon footprint of consumer products will be an important expectation of a majority of next-generation consumers. The deployment of green technologies is

also essential for renewable feedstocks in order to fulfill their promise for producing sustainable fuels and chemicals. Renewable feedstocks cannot yet be processed with existing technologies because their chemical structures and compositions (their relative C, H, and O contents) are varied and diverse compared to conventional crude oil. The fledgling biorefinery industry is uniquely positioned to accept new technology concepts. Chemical catalysis is a promising avenue to develop atom-economical and energy-efficient biomass conversion technologies. To address this challenge, multidisciplinary expertise is needed for the design and synthesis of novel metal-based catalysts, high-pressure chemistry, solvent engineering, multiphase reaction engineering, and multiscale modeling. Specifically, the various system elements (i.e., catalysts, solvents, reactors, and separators) should be designed such that when integrated, the resulting system displays enhanced rates and selectivity with reduced environmental footprint.

Acknowledgments

Much of the author's work described in this entry was made possible by NSF ERC Grant EEC-0310689, the Kansas Technology Enterprise Corporation, and the support of the University of Kansas through the Dan F. Servey Distinguished Professorship.

Bibliography

Primary Literature

1. Jessop PG, Subramaniam B (2007) Gas-expanded liquids. *Chem Rev* 107:2666–2694
2. Lopez-Castillo ZK, Aki SNVK, Stadtherr MA, Brennecke JF (2006) Enhanced solubility of oxygen and carbon monoxide in CO_2 -expanded liquids. *Ind Eng Chem Res* 45:5351–5360
3. Lopez-Castillo ZK, Aki SNVK, Stadtherr MA, Brennecke JF (2008) Enhanced solubility of hydrogen in CO_2 -expanded liquids. *Ind Eng Chem Res* 47:570–576
4. Zevnik L, Levec J (2007) Hydrogen solubility in CO_2 -expanded 2-propanol and in propane-expanded 2-propanol determined by an acoustic sensor. *J Supercrit Fluids* 41:335–342
5. Xie Z, Snavely WK, Scurto AM, Subramaniam B (2009) Solubilities of CO and H_2 in neat and CO_2 -expanded hydroformylation reaction mixtures containing 1-octene and nonanal up to 353 K and 9 Mpa. *J Chem Eng Data* 54:1633–1642
6. Sheldon RA (1994) Consider the environmental quotient. *Chem Tech* 24:38–47

7. Sheldon RA, Arends IWCE, Hanefeld U (2007) Green chemistry and catalysis. Wiley, Weinheim
8. Tundo AP, Black DS, Breen J, Collins T, Memoli S, Miyamoto J, Poliakov M, Tumas W (2000) Synthetic pathways and processes in green chemistry: introductory overview. *Pure Appl Chem* 72:1207–1228
9. DeSimone JM (2002) Practical approaches to green solvents. *Science* 297:799–803
10. Adams DJ, Dyson PJ, Tavener SJ (2004) Chemistry in alternative reaction media. Wiley, Chichester
11. Eckert CA, Liotta CL, Bush B, Brown JS, Hallett JP (2004) Sustainable reactions in tunable solvents. *J Phys Chem B* 108:18108–18118
12. Seki T, Baiker A (2009) Catalytic oxidations in dense carbon dioxide. *Chem Rev* 109:2409–2454
13. Morgenstern DA, LeLacheur RM, Morita DK, Borkowsky SL, Feng S, Brown GH, Luan L, Gross MF, Burk MJ, Tumas W (1996) Supercritical carbon dioxide as a substitute solvent for chemical synthesis and catalysis. In: Anastas PT, Williamson TC (eds) Green chemistry: designing chemistry for the environment, ACS symposium series vol 626. American Chemical Society, Washington, DC, pp 132–151
14. Jessop PG, Leitner W (1999) Chemical synthesis using supercritical fluids. Wiley, Weinheim
15. Amandi R, Hyde J, Poliakov M (2003) Heterogeneous reactions in supercritical carbon dioxide. In: Aresta M (ed) Carbon dioxide recovery and utilization. Kluwer, Dordrecht, pp 169–180
16. DeSimone JM, Tumas W (2003) Green chemistry using liquid and supercritical carbon dioxide. Oxford University Press, New York
17. Gordon CM, Leitner W (2004) Supercritical fluids as replacements for conventional organic solvents. *Chim Oggi* 22:39–41
18. Beckman EJ (2002) Using CO₂ to produce chemical products sustainably. *Environ Sci Technol* 36:347A–353A
19. Licence P, Poliakov M (2005) Economics and scale-up. In: Cornils B, Hermann WA, Horváth IT, Leitner W, Mecking S, Olivier-Bourbigou H, Vogt D (eds) Multiphase homogeneous catalysis, vol 2. Wiley, Weinheim, pp 734–746
20. Arai M, Fujita SI, Shirai M (2009) Multiphase catalytic reaction in/under dense phase CO₂. *J Supercrit Fluids* 47:351–356
21. Li CJ, Chan TH (1997) Organic reactions in aqueous media. Wiley, New York
22. Cornils B, Herrmann WA (1998) Aqueous-phase organometallic catalysis. Wiley, Weinheim
23. Savage PE (2009) A perspective on catalysis in sub- and supercritical water. *J Supercrit Fluids* 47:407–414
24. Musie G, Wei M, Subramaniam B, Busch DH (2001) Catalytic oxidations in carbon dioxide-based reaction media, including novel CO₂-expanded phases. *Coord Chem Rev* 219–221: 789–820
25. Hutchenson KW, Scurto AM, Subramaniam B (2009) Gas-expanded liquids and near-critical media: green chemistry and engineering, vol 1006, ACS symposium series. American Chemical Society, Washington, DC
26. Akién GR, Poliakov M (2009) A critical look at reactions in class I and II gas-expanded liquids using CO₂ and other gases. *Green Chem* 11:1083–1100
27. Scurto AM, Hutchenson KW, Subramaniam B (2009) Gas-expanded liquids (GXLs): fundamentals and applications. In: Hutchenson KW, Scurto AM, Subramaniam B (eds) Gas-expanded liquids and near-critical media: green chemistry and engineering, vol 1006, ACS symposium series. American Chemical Society, Washington, DC, pp 3–37
28. Wasserscheid P, Welton T (2002) Ionic liquids in synthesis. Wiley, Weinheim
29. Rogers RD, Seddon KR, Volkov S (2003) Green industrial applications of ionic liquids. Kluwer, Dordrecht
30. Părvulescu VI, Hardacre C (2007) Catalysis in ionic liquids. *Chem Rev* 107:2615–2665
31. Jessop PG, Heldebrant DJ, Xiaowang L, Eckert CA, Liotta CL (2005) Reversible nonpolar-to-polar solvent. *Nature* 436:1102
32. Liu Y, Jessop PG, Cunningham M, Eckert CA (2006) Liotta CL, switchable surfactants. *Science* 313:958–960
33. Phan CD, Heldebrant DJ, Huttenhower H, John E, Li X, Pollet P, Wang R, Eckert CA, Liotta CL, Jessop PG (2008) Switchable solvents consisting of amidine/alcohol or guanidine/alcohol mixtures. *Ind Eng Chem Res* 47:539–545
34. Phan L, Brown H, White J, Hodgson A, Jessop PG (2009) Soybean oil extraction and separation using switchable or expanded solvents. *Green Chem* 11:53–59
35. Phan L, Jessop PG (2009) Switching the hydrophilicity of a solute. *Green Chem* 11:307–308
36. Aghosseini A, Ren W, Scurto AM (2009) Understanding biphasic ionic liquid/CO₂ systems for homogeneous catalysis: hydroformylation. *Ind Eng Chem Res* 48:4254–4265
37. Anastas P, Warner JC (1998) Green chemistry: theory and practice. Oxford University Press, New York
38. Anastas PT, Zimmerman JB (2003) Design through the 12 principles of green engineering. *J Environ Sci Technol* 37:95A–101A
39. Allen DT, Shonnard DR (2001) Green engineering: environmentally conscious design of chemical processes. Prentice Hall, New York
40. Dudukovic MP (2009) Frontiers in reactor engineering. *Science* 325:698–701
41. Kordikowski A, Schenk AP, Van Nielen RM, Peters CJ (1995) Volume expansions and vapor-liquid equilibria of binary mixtures of a variety of polar solvents and certain near-critical solvents. *J Supercrit Fluids* 8:205–216
42. Ren W, Scurto AM (2007) High-Pressure phase equilibria with compressed gases. *Rev Sci Instrum* 78:125104–125107
43. Heldebrant DJ, Witt H, Walsh S, Ellis T, Rauscher J, Jessop PG (2006) Liquid polymers as solvents for catalytic reductions. *Green Chem* 8:807–815
44. Ren W, Sensenich B, Scurto AM (2010) High-pressure phase equilibria of carbon dioxide (CO₂) + n-alkyl-imidazolium bis (trifluoromethylsulfonyle)amide ionic liquids. *J Chem Thermodyn* 42:305–311

45. Rajagopalan B, Wie M, Musie GT, Subramaniam B, Busch DH (2003) Homogeneous catalytic epoxidation of organic substrates in CO₂-expanded solvents in the presence of water soluble oxidants and catalysts. *Ind Eng Chem Res* 42:6505–6510
46. Ohgaki K, Takata H, Washida T, Katayama T (1988) Phase equilibria of four binary systems containing propylene. *Fluid Phase Equilib* 43:105–113
47. Peng DB, Robinson DT (1976) A new two-constant equation of state. *Ind Eng Chem Fund* 15:59–64
48. Houndonougbo Y, Jin H, Rajagopalan B, Wong K, Kuczera K, Subramaniam B, Laird BB (2006) Phase equilibria in carbon dioxide-expanded solvents: experiment and molecular simulations. *J Phys Chem B* 110:13195–13202
49. Swalina C, Arzhantsev S, Li HP, Maroncelli M (2008) Solvation and solvatochromism in CO₂-expanded liquids. 3. the dynamics of nonspecific preferential solvation. *J Phys Chem B* 112:14959–14970
50. Subramaniam B (2010) Gas-expanded liquids for sustainable catalysis and novel materials. *Coord Chem Rev* 254:1843–1853
51. Sih R, Dehghani F, Foster NR (2007) Viscosity measurements on gas expanded liquid systems-methanol and carbon dioxide. *J Supercrit Fluids* 41:148–157
52. Kelkar MS, Maginn EJ (2007) Effect of temperature and water content on the shear viscosity of the ionic liquid 1-ethyl-3-methylimidazolium bis(trifluoromethanesulfonyl)imide as studied by atomistic simulations. *J Phys Chem B* 111:4867–4876
53. Ahosseini A, Ortega E, Sensenich B, Scurto AM (2009) Viscosity of n-alkyl-3-methyl-imidazolium bis(trifluoromethylsulfonyl) amide ionic liquids saturated with compressed CO₂. *Fluid Phase Equilib* 286:62–68
54. Maxey NB (2006) Transport and phase-transfer catalysis in gas-expanded liquids. PhD Dissertation, Georgia Institute of Technology, Atlanta
55. Lin IH, Tan CS (2008) Diffusion of benzonitrile in CO₂-expanded ethanol. *J Chem Eng Data* 53:1886–1891
56. Roškar V, Dombro RA, Prentice GA, Westgate CR, McHugh MA (1992) Comparison of the dielectric behavior of mixtures of methanol with carbon dioxide and ethane in the mixture-critical and liquid regions. *Fluid Phase Equilib* 77:241–259
57. Wyatt VT, Bush D, Lu J, Hallett JP, Liotta CL, Eckert CA (2005) Determination of solvatochromic solvent parameters for the characterization of gas-expanded liquids. *J Supercrit Fluids* 36:16–22
58. Abbott AP, Hope EG, Mistry R, Stuart AM (2009) Probing the structure of gas expanded liquids using relative permittivity, density and polarity measurements. *Green Chem* 11:1530–1535
59. Ford JW, Janakat ME, Liu J, Liotta CL, Eckert CA (2008) Local polarity in CO₂-expanded acetonitrile: A nucleophilic substitution reaction and solvatochromic probes. *J Org Chem* 73:3364–3368
60. Seki TJ, Grunwaldt JD, Baiker A (2009) In situ attenuated total reflection infrared spectroscopy of imidazolium-based room-temperature ionic liquids under “supercritical” CO₂. *J Phys Chem B* 113:114–122
61. Burgi T, Baiker A (2006) Attenuated total reflection infrared spectroscopy of solid catalysts functioning in the presence of liquid-phase reactants. *Adv Catal* 50:227–283
62. Guha D, Jin H, Dudukovic MP, Ramachandran PA, Subramaniam B (2007) Mass transfer effects during homogeneous 1-octene hydroformylation in CO₂-expanded solvent: modeling and experiments. *Chem Eng Sci* 62:4967–4975
63. Lyon CJ, Subramaniam B, Pereira CJ (2001) Enhanced isooctane yields for 1-butene/isobutane alkylation on SiO₂-supported Nafion® in supercritical carbon dioxide. In: Spivey JJ, Roberts GW, Davis BH (eds) *Catalyst deactivation 2001. Studies in surface science and catalysis*, vol 139. Elsevier, Amsterdam, pp 221–228
64. (a) Webb PB, Kunene TE, Cole-Hamilton DJ (2005) Continuous flow homogeneous hydroformylation of alkenes using supercritical fluids. *Green Chem* 7:373–379; (b) Sellin MF, Webb PB, Cole-Hamilton DJ (2001) *Chem Commun* 781
65. Thomas CA, Bonilla RJ, Huang Y, Jessop PG (2001) Hydrogenation of carbon dioxide catalysed by ruthenium trimethylphosphine complexes: effect of gas pressure and additives on rate in the liquid phase. *Can J Chem* 79:719–724
66. Combes GB, Dehghani F, Lucien FP, Dillow AK, Foster NR (2000) Asymmetric catalytic hydrogenation in CO₂ expanded methanol-an application of gas anti-solvent reaction (GASR). In: Abraham MA, Hesketh RP (eds) *Reaction engineering for pollution prevention*. Elsevier, Amsterdam, pp 173–181
67. Combes G, Coen E, Dehghani F, Foster NR (2005) Dense CO₂ expanded methanol solvent system for synthesis of naproxen via enantioselective hydrogenation. *J Supercrit Fluids* 36:127–136
68. Floris T, Kluson P, Muldoon MJ, Pelantova H (2010) Notes on the asymmetric hydrogenation of methyl acetoacetate in neoteric solvents. *Cat Lett* 134:279–287
69. Solinas M, Pfaltz A, Cozzi P, Leitner W (2004) Enantioselective hydrogenation of imines in ionic liquid/carbon dioxide media. *J Am Chem Soc* 126:16142–16147
70. Jessop PG, Stanley R, Brown RA, Eckert CA, Liotta CL, Ngo TT, Pollet P (2003) Neoteric solvents for asymmetric hydrogenation: supercritical fluids, ionic liquids, and expanded ionic liquids. *Green Chem* 5:123–128
71. Jessop PG, DeHaai S, Wynne DC, Nakawata D (2000) Carbon dioxide gas accelerates solventless synthesis. *Chem Commun* 8:693–694
72. Scurto AM, Leitner W (2006) Melting point depression of organic ionic solids/liquids with carbon dioxide for enhanced catalytic processes. *Chem Commun* 3681–3683
73. Ahosseini A, Ren W, Scurto AM (2009) Hydrogenation in biphasic ionic liquid/CO₂ systems. In: Hutchenson KW, Scurto AM, Subramaniam B (eds) *Gas expanded liquids and near-critical media: green chemistry and engineering*, vol 1006, ACS symposium series. American Chemical Society, Washington, DC, pp 218–234

74. Jin H, Subramaniam B (2004) Catalytic hydroformylation of 1-octene in CO₂-expanded solvent media. *Chem Eng Sci* 59:4887–4893
75. Jin H, Subramaniam B, Ghosh A, Tunge J (2006) Intensification of catalytic olefin hydroformylation in CO₂-expanded media. *AIChE J* 52:2575–2591
76. Koeken ACJ, Benes NE, van den Broeke LJP, Keurentjes JTF (2009) Efficient hydroformylation in dense carbon dioxide using phosphorus ligands without perfluoroalkyl substituents. *Adv Syn Catal* 351:1142–1450
77. Hemminger O, Marteel A, Mason MR, Davies JA, Tadd AR, Abraham MA (2002) Hydroformylation of 1-hexene in supercritical carbon dioxide using a heterogeneous rhodium catalyst. 3. Evaluation of solvent effects. *Green Chem* 4: 507–512
78. Webb PB, Sellin MF, Kunene TE, Williamson S, Slawin AMZ, Cole-Hamilton DJ (2003) Continuous flow hydroformylation of alkenes in supercritical fluid-ionic liquid biphasic systems. *J Am Chem Soc* 125:15577–15588
79. Frisch AC, Webb PB, Zhao G, Muldoon MJ, Pogorzelec PJ, Cole-Hamilton DJ (2007) Solventless continuous flow homogeneous hydroformylation of 1-octene. *Dalton Trans* 47:5531–5538
80. Wei M, Musie GT, Busch DH, Subramaniam B (2002) CO₂-expanded solvents: unique and versatile media for performing homogeneous catalytic oxidations. *J Am Chem Soc* 124: 2513–2517
81. Wei M, Musie GT, Busch DH, Subramaniam B (2004) Autoxidation of 2,6-di-tertbutylphenol with cobalt Schiff base catalysts by oxygen in CO₂-expanded liquids. *Green Chem* 6:387–393
82. Zuo X, Niu F, Snavely WK, Subramaniam B, Busch DH (2010) Liquid phase oxidation of *p*-xylene to terephthalic acid at medium-high temperatures: multiple benefits of CO₂-expanded liquids. *Green Chem* 12:260–267
83. Trent DT (1996) Propylene oxide Kirk-Othmer encyclopedia of chemical technology, vol 20, 4th edn. Wiley, New York, pp 271–302
84. Thiele GF, Roland E (1997) Propylene epoxidation with hydrogen peroxide and titanium silicalite catalyst: activity, deactivation and regeneration of the catalyst. *J Mol Catal A Chem* 117:351–356
85. Danciu T, Beckman EJ, Hancu D, Cochran RN, Grey R, Hajnik DM, Jewson J (2002) Direct synthesis of propylene oxide with CO₂ as the solvent. *Angew Chem Int Ed* 42:1140–1142
86. Laufer W, Meiers R, Holderich W (1999) Propylene epoxidation with hydrogen peroxide over palladium containing titanium silicalite. *J Mol Catal A Chem* 141:215–221
87. Jenzer G, Mallat T, Maciejewski M, Eigenmann F, Baiker A (2001) Continuous epoxidation of propylene with oxygen and hydrogen on a Pd-Pt/TS-1 catalyst. *Appl Catal A* 208:125–133
88. Nolen SA, Lu J, Brown JS, Pollet P, Eason BC, Griffith KN, Glaser R, Bush D, Lamb DR, Liotta CL, Eckert CA, Thiele GF, Bartels KA (2002) Olefin epoxidations using supercritical carbon dioxide and hydrogen peroxide without added metallic catalysts or peroxy acids. *Ind Eng Chem Res* 41:316–323
89. Hancu D, Green H, Beckman EJ (2002) H₂O₂ in CO₂/H₂O biphasic systems: green synthesis and epoxidation reactions. *Ind Eng Chem Res* 41:4466–4474
90. Zha YJ, Zhang JL, Han BX, Hu SQ, Li W (2010) CO₂-controlled reactors: epoxidation in emulsions with droplet size from micron to nanometre scale. *Green Chem* 12:452–457
91. Herrmann WA, Fischer RW, Marz DW (1991) Methyltrioxorhenium as catalyst for olefin metathesis. *Angew Chem Int Ed Engl* 30:1636–1638
92. Rudolph J, Reddy KL, Chiang JP, Sharpless KB (1997) Highly efficient epoxidation of olefins using aqueous H₂O₂ and catalytic methyltrioxorhenium/pyridine: pyridine-mediated ligand acceleration. *J Am Chem Soc* 119:6189–6190
93. Wang WD, Espenson JH (1998) Effects of pyridine and its derivatives on the equilibria and kinetics pertaining to epoxidation reactions catalyzed by methyltrioxorhenium. *J Am Chem Soc* 120:11335–11341
94. Yin G, Busch DH (2009) Mechanistic details to facilitate applications of an exceptional catalyst, methyltrioxorhenium: encouraging results from oxygen-18 isotopic probes. *Catal Lett* 130:52–55
95. Lee HJ, Shi TP, Busch DH, Subramaniam B (2007) A greener, pressure intensified propylene epoxidation process with facile product separation. *Chem Eng Sci* 62:7282–7289
96. Lee HJ, Ghanta M, Busch DH, Subramaniam B (2010) Towards a CO₂-free ethylene oxide process: homogeneous ethylene epoxidation in gas-expanded liquids. *Chem Eng Sci* 65: 128–134
97. Azarnoosh A, Mcketta JJ (1959) Solubility of propylene in water. *J Chem Eng Data* 4:211–212
98. Yorzane M, Sadamoto S, Yoshimura S (1968) Low-temperature vapor-liquid equilibria. Nitrogen-propylene and carbon dioxide-methane systems. *Kagaku Kogaku Ronbun* 32:257–264
99. Lee HJ, Shi TP, Subramaniam B, Busch DH (2006) Selective oxidation of propylene to propylene oxide in CO₂ expanded liquid system. In: Schmidt SR (ed) *Catalysis of organic reactions*. CRC Press, Boca Raton, pp 447–451
100. Weissmermel K (2003) *Industrial organic chemistry*, 4th edn. Wiley, Weinheim, pp 145–153
101. Haneda A, Seki T, Kodama D, Kato M (2006) High-pressure phase equilibrium for ethylene + methanol at 278.15 K and 283.65 K. *J Chem Eng Data* 51:268–271
102. West KN, Wheeler C, McCauley JP, Griffith KN, Bush D, Liotta CL, Eckert CA (2001) In situ formation of alkylcarbonic acids with CO₂. *J Phys Chem A* 105:3947–3948
103. Chamblee TS, Weikel RR, Nolen SA, Liotta CL, Eckert CA (2004) Reversible *in situ* acid formation for -pinene hydrolysis using CO₂ expanded liquid and hot water. *Green Chem* 6:382–386
104. Gohres JL, Marin AT, Lu J, Liotta CL, Eckert CA (2009) Spectroscopic investigation of alkylcarbonic acid formation and dissociation in CO₂-expanded alcohols. *Ind Eng Chem Res* 48:1302–1306

105. Alemán PA, Boix C, Poliakoff M (1999) Hydrolysis and saponification of methyl benzoates. *Green Chem* 1:65–68
106. Hunter SE, Savage PE (2004) Recent advances in acid- and base-catalyzed organic synthesis in high-temperature liquid water. *Chem Eng Sci* 59:4903–4909
107. Throckmorton PE, Hansen LI, Christensen RC, Pryde EH (1968) Laboratory optimization of process variables in reductive ozonolysis of methyl soyate. *J Am Oil Chem Soc* 45:59–62
108. Nickell EC, Albi M, Privett OS (1976) Ozonization products of unsaturated fatty acid methyl esters. *Chem Phys Lipids* 17:378–388
109. Nishikawa N, Yamada K, Matsutani S, Higo M, Kigawa H, Inagaki T (1995) Structures of ozonolysis products of methyl oleate obtained in a carboxylic acid medium. *J Am Oil Chem Soc* 72:735–740
110. O'Brien M, Baxendale IR, Ley SV (2010) Flow ozonolysis using a semipermeable Teflon AF-2400 membrane to effect gas-liquid contact. *Org Lett* 12:1596–1598
111. Subramaniam B, Busch DH, Danby A, Binder TP (2008) Ozonolysis reactions in liquid CO₂ and CO₂-expanded solvents. U. S. Patent Application, 20090118498
112. Del Moral D, Osuna AMB, Cordoba A, Moreto JM, Veciana J, Ricart S, Ventosa N (2009) Versatile chemoselectivity in Ni-catalyzed multiple bond carbonylations and cyclocarbonylations in CO₂-expanded liquids. *Chem Commun* 31:4723–4725
113. Zwolak G, Jayasinghe NS, Lucien FP (2006) Catalytic chain transfer polymerisation of CO₂-expanded methyl methacrylate. *J Supercrit Fluids* 38:420–426
114. Wyatt VT, Haas MJ (2009) Production of fatty acid methyl esters via the in situ transesterification of soybean oil in carbon dioxide-expanded methanol. *J Am Oil Chem Soc* 86:1009–1016
115. George J, Patel Y, Pillai SM, Munshi P (2009) Methanol assisted selective formation of 1, 2-glycerol carbonate from glycerol and carbon dioxide using (Bu₂SnO)-Bu-*n* as a catalyst. *J Mol Catal A Chem* 304:1–7
116. Baiker A (1999) Supercritical fluids in heterogeneous catalysis. *Chem Rev* 99:453–474
117. Grunwaldt JD, Wandeler R, Baiker A (2003) Supercritical fluids in catalysis: opportunities of in situ spectroscopic studies and monitoring phase behavior. *Catal Rev Sci Eng* 45:1–96
118. Beckman EJ (2004) Supercritical and near-critical CO₂ in green chemical synthesis and processing. *J Supercrit Fluids* 28: 121–191
119. Devetta L, Giovanzana A, Canu P, Bertuccio A, Minder B (1999) Kinetic experiments and modeling of a three-phase catalytic hydrogenation reaction in supercritical CO₂. *Catal Today* 48:337–345
120. Chouchi D, Gourguillon D, Courel M, Vital J, Nunes da Ponte M (2001) The influence of phase behavior on reactions at supercritical conditions: the hydrogenation of α -pinene. *Ind Eng Chem Res* 40:2551–2554
121. Xu D, Carbonell RG, Kiserow DJ, Roberts GW (2005) Hydrogenation of polystyrene in CO₂-expanded solvents: catalyst poisoning. *Ind Eng Chem Res* 44:6164–6170
122. Chan JC, Tan CS (2006) Hydrogenation of tetralin over Pt/ γ -Al₂O₃ in trickle-bed reactor in the presence of compressed CO₂. *Energy Fuels* 20:771–777
123. Phiong H-S, Cooper CG, Adesina AA, Lucien FP (2008) Kinetic modelling of the catalytic hydrogenation of CO₂-expanded alpha-methylstyrene. *J Supercrit Fluids* 46:40–46
124. Xie X, Liotta CL, Eckert CA (2004) CO₂-protected amine formation from nitrile and imine hydrogenation in gas-expanded liquids. *Ind Eng Chem Res* 43:7907–7911
125. Jenzer G, Schneider MS, Wandeler R, Mallat T, Baiker A (2001) Palladium-catalyzed oxidation of octyl alcohols in “supercritical” carbon dioxide. *J Catal* 199:141–148
126. Kerler B, Robinson RE, Borovik AS, Subramaniam B (2004) Application of CO₂-expanded solvents in heterogeneous catalysis: a case study. *Appl Catal B* 49:91–98
127. Sharma S, Kerler B, Subramaniam B, Borovik AS (2006) Immobilized metal complexes in porous hosts: catalytic oxidation of substituted phenols in CO₂ media. *Green Chem* 8:972–977
128. Stobrawe A, Makarczyk P, Maillet C, Muller JL, Leitner W (2008) Solid-phase organic synthesis in the presence of compressed carbon dioxide. *Angew Chem Int Ed* 47:6674–6677
129. Sarsani VSR, Lyon CJ, Hutchenson KW, Harmer MA, Subramaniam B (2007) Continuous acylation of anisole by acetic anhydride in mesoporous solid acid catalysts: reaction media effects on catalyst deactivation. *J Catal* 245:184–190
130. Chateaufneuf JE, Nie K (2000) An investigation of a Friedel-Crafts alkylation reaction in homogeneous supercritical CO₂ and under subcritical and splitphase reaction conditions. *Adv Environ Res* 4:307–312
131. Garton RD, Ritchie JT, Caers RE (2003) Oxo process. PCT International Application, WO 2003/082789 A2
132. Bhanage BM, Divekar SS, Deshpande RM, Chaudhari RV (1997) Kinetics of hydroformylation of 1-dodecene using homogeneous HRh(CO)(PPh₃)₃ catalyst. *J Mol Catal A Chem* 115:247–257
133. Purwanto P, Deshpande RM, Delmas H, Chaudhari RV (1996) Solubility of hydrogen, carbon monoxide, and 1-octene in various solvents and solvent mixtures. *J Chem Eng Data* 41:1414–1417
134. Guha D, Jin H, Dudukovic MP, Ramachandran PA, Subramaniam B (2007) Mass transfer effects during homogeneous 1-octene hydroformylation in CO₂-expanded solvent: modeling and experiments. *Chem Eng Sci* 62:4967–4975
135. Fang J, Jin H, Ruddy T, Pennybaker K, Fahey D, Subramaniam B (2007) Economic and environmental impact analyses of catalytic olefin hydroformylation in CO₂-expanded liquid (CXL) media. *Ind Eng Chem Res* 46:8687–8692
136. Jana R, Tunge JA (2009) A homogeneous, recyclable rhodium (I) catalyst for the hydroarylation of Michael acceptors. *Org Lett* 11:971–974

137. Wang R, Cai F, Jin H, Xie Z, Subramaniam B, Tunge JA (2009) Hydroformylation in CO₂-expanded media. In: Hutchenson KW, Scurto AM, Subramaniam B (eds) Gas-expanded liquids and near-critical media: green chemistry and engineering, vol 1006, ACS symposium series. American Chemical Society, Washington, DC, pp 202–217
138. Fang J, Jana R, Tunge JA, Subramaniam B (2011) Continuous homogeneous hydroformylation with bulky rhodium catalyst complexes retained by nano-filtration membranes. *Appl Catal A: Gen* 393:294–301
139. Subramaniam B, Tunge JA, Jin H, Ghosh A (2008) Tuning product selectivity in catalytic hydroformylation reactions with CO₂-expanded liquids. US Patent 7.365,234, 29 Apr 2008
140. Horstmann S, Grybat A, Kato R (2004) Experimental determination and prediction of gas solubility data for oxygen in acetonitrile. *J Chem Thermodyn* 36:1015–1018

Books and Reviews

- Anastas P, Eghbali N (2010) Green chemistry: principles and practice. *Chem Soc Rev* 39:301–312
- McHugh MA, Krukonis VJ (1994) Supercritical fluid extraction: principles & practice. Butterworth-Heinemann, Boston
- Muldoon MJ (2010) Modern multiphase catalysis: new developments in the separation of homogeneous catalysts. *Dalton Trans* 39:337–348
- Olivier-Bourbigou H, Magna L, Morvan D (2010) Ionic liquids and catalysis: recent progress from knowledge to applications. *Appl Cat A* 373:1–56

Gas to Liquid Technologies

MARIANNA ASARO¹, RONALD M. SMITH²

¹SRI International, Menlo Park, CA, USA

²SRI Consulting, Menlo Park, CA, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Survey of Specific Gas to Liquids Technologies
 Liquefied Natural Gas (LNG)
 Methane Reforming in F-T GTL
 Methane Reforming for Methanol/DME Synthesis
 Fischer–Tropsch Synthesis
 F-T Product Upgrading

Methanol
 Dimethylether
 Methanol to Gasoline and/or Diesel
 Methanol to Chemicals
 Ammonia
 Direct Conversion of Natural Gas to Liquids
 Future Directions
 Bibliography

Glossary

Autothermal reforming (ATR) The reaction of oxygen and carbon dioxide or steam with methane to form synthesis gas, wherein the exothermic partial oxidation of methane provides energy for the endothermic steam reforming of methane. ATR is also used in reference to the autothermal reformer itself.

bbl Barrels (of oil).

BOE Barrel of oil equivalent.

Btu Also btu, British thermal unit, a measure of energy content.

Catalytic membrane reactor (CMR) A flow-through reactor used to influence an equilibrium-limited reaction to proceed further in the forward direction via selective transport of reactant(s) or product(s) across the membrane.

CPOx Catalytic partial oxidation.

CTL Coal-to-liquids.

DME Dimethylether.

DOE US Department of Energy.

Fischer–Tropsch reaction The catalytic conversion of synthesis gas to primarily hydrocarbons, the discovery being credited to Franz Fischer and Hans Tropsch.

Gas-Heated reforming (GHR) Use of heat available by recycling process gas (tail gas) downstream of an ATR for steam methane reforming, in a heat exchanger type reactor.

Gas to liquids (GTL) The conversion of gas to liquid fuels and/or chemicals.

Heat exchange reforming (HER) See GHR.

kWh Kilowatt hours, a measure of energy.

Light distillate A distillation cut of low molecular weight and low boiling range, obtained from refining of hydrocarbonaceous feedstocks, used to produce liquefied petroleum gas (LPG), gasoline, and naphtha.

Liquefied natural gas (LNG) Natural gas that has been converted to liquid form for transport or storage.

Middle distillate A distillation cut of mid-range boiling point, obtained from refining hydrocarbonaceous feedstocks, containing hydrocarbons ranging from C_5 through about C_{20} or C_{22} . When further distilled, the portion of middle distillates containing C_5 through about C_{15} is often referred to as naphtha, and the portion containing C_{16} through up to C_{22} is referred to as diesel. The naphtha is often distilled further to produce gasoline and kerosene/jet fuel, or can be used as feed for a naphtha cracker unit to make light olefins. (Less commonly, the gasoline cut is initially collected along with the light distillates.)

Natural gas liquids (NGL) The purified and condensed portion of natural gas consisting of gaseous hydrocarbons heavier than methane specifically ethane (C_2H_6), propane (C_3H_8), *n*-butane ($n-C_4H_{10}$), and isobutane (*i*- C_4H_{10}).

Partial oxidation (POx) The controlled oxidation of natural gas (primarily CH_4) with oxygen (O_2) such that syngas is formed, rather than forming carbon dioxide (CO_2) via complete combustion.

Pre-reforming The use of an adiabatic, preheating zone upstream of an ATR reactor for the purpose of catalytically converting C_{2+} hydrocarbons to a mixture of methane (CH_4), hydrogen (H_2), carbon monoxide (CO), and carbon dioxide (CO_2), thereby allowing oxygen (O_2) to be used more efficiently in the reformer.

scf Standard cubic feet, a measure of gas volume.

Steam methane reforming (SMR) The high-temperature catalytic reaction of steam with methane to give synthesis gas.

Synthesis gas (syngas) A mixture of primarily hydrogen and carbon monoxide produced by gasification or reforming of hydrocarbonaceous materials, used to synthesize fuels or chemicals.

Water gas shift reaction (WGSR) The gas phase reaction of carbon monoxide (CO) with water to form carbon dioxide (CO_2) and hydrogen (H_2).

Definition of the Subject

Like oil and coal, natural gas is not what first comes to mind when considering sustainable fuel sources.

Yet as for other fossil sources, conversion of natural gas to transportation fuel is currently more affordable than conversion of renewable resources such as wind and solar, which are technically far away from availability at even a fraction of the scale required to have significant impact on meeting global demand over the coming decades. Given that global proved natural gas reserves are currently estimated as capable of producing more than 1,100 billion equivalent barrels of oil (the energy equivalent of 42 cubic miles of oil) [1], natural gas is a key contributor when considering a sustainable global fuel supply.

The term “gas to liquids” (GTL) is frequently used in reference to the chemical transformation of natural gas to liquid fuels via the Fischer–Tropsch (F-T) technology. In broader usage, the term “GTL” refers to the transformation of natural gas into any liquid, including other fuels, such as liquefied natural gas (LNG), methanol, dimethylether (DME), methyl-*tert*-butyl ether (MTBE), and chemicals such as ammonia (itself an important feedstock, particularly in the fertilizer industry), light olefins, and methanol and dimethylether intended for chemical use.

For clarity, the term “GTL” is used herein to refer to the chemical conversion of natural gas to liquids. However, a practical discussion of GTL should also include LNG, which is produced from natural gas by a phase change rather than a chemical transformation. Strategists in national governments and energy companies alike weigh all of these options when considering how best to distribute and monetize natural gas.

Modern GTL synthesis technology can produce synthetic fuels (synfuels) that burn more cleanly than conventional fuels derived from petroleum. F-T GTL technology can produce a virtually sulfur-free diesel fuel, much cleaner than conventional diesel – reducing smog and acid rain – and can also produce gasoline, jet fuel, or chemicals. GTL performed at or near the gas well also represents an emerging option to transport large quantities of transportation fuels and chemical feedstocks from natural gas, using commercially available tankers.

Production of liquid fuels and chemicals are lower volume applications compared to use of natural gas for power or as a fuel source for heating. Commercial

application of GTL might increase in magnitude as the ease of oil recovery decreases and its price increases over the next few decades.

Introduction

The methane content of natural gas varies greatly within the range of about 50–99%, typically about 95%. The remainder is a multicomponent mixture including the heavier gaseous hydrocarbons ethane, propane, butane, pentane, and some higher molecular weight hydrocarbons; the acid gases carbon dioxide, hydrogen sulfide; mercaptans such as methanethiol and ethanethiol; water, nitrogen gas; helium gas; liquid hydrocarbons; and trace mercury. Thus the preponderance of species in natural gas are high in energy content, with cleanup of contaminants necessary but greatly simplified compared to that required when using coal. Likewise the CO₂ emissions from combustion of natural gas [2] are considerably less than from oil [3] or coal [4]: 0.40 lb/kWh compared to 0.61 and 0.79 lb/kWh, respectively.

The bulk of natural gas used today goes directly into power generation (32%), residential and commercial heating (27%), and industrial use (22%). Only a small portion, around 5%, is used as a chemical feedstock, and therefore chemical uses have minimal impact on supply and pricing decisions other than in stranded gas regions where the gas is geographically remote from the main end-user markets. Thus in the Middle East, for example, which has large reserves but relatively small local demand, the proportion of natural gas going into feedstocks is higher (around 27%).

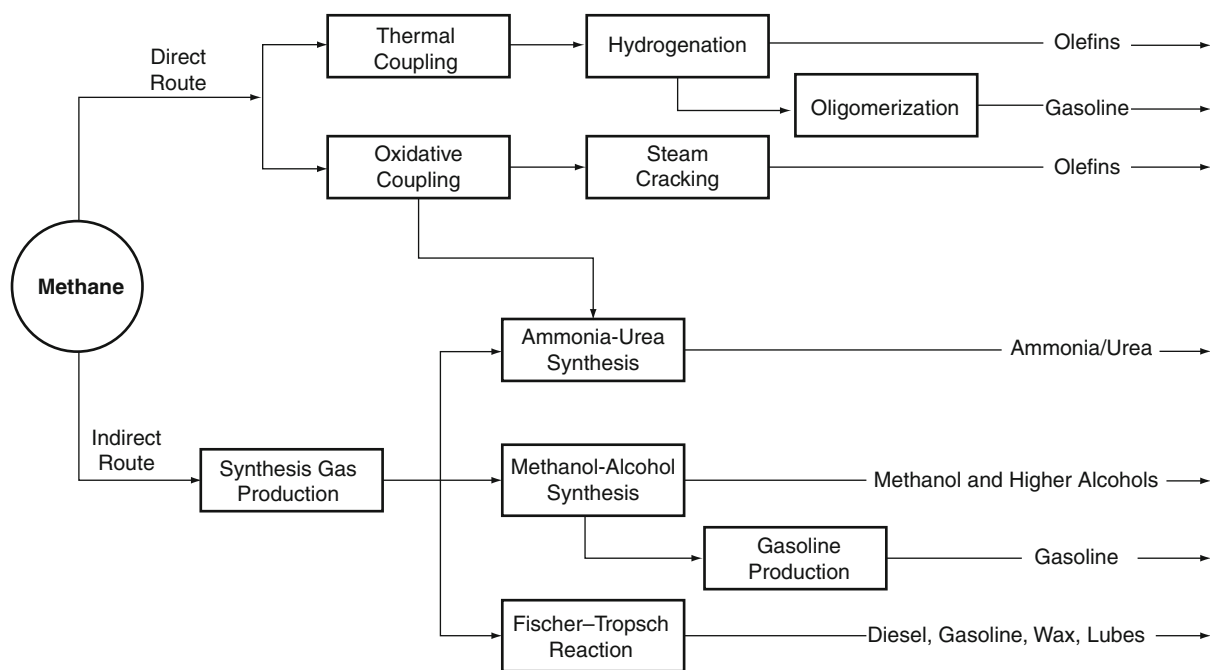
Location and transportation are key factors in defining the need for, and choice of, GTL technology. About 40% of known natural gas reserves are of the stranded type [5], including remote gas (located far from existing pipelines) and deeply buried or otherwise inaccessible gas. Remote gas alone accounts for about 16% of the world's proved natural gas reserves [6]. Related factors determine the type of transportation used, and therefore, the type of GTL processing used, including the size of the reserve, how much time is available to monetize the natural gas, the distance to market, and the gas processing requirements. Environmental influences include a growing resistance to the conventional practice of flaring (or reinjecting) gas that otherwise

would be released in association with oil recovery operations (i.e., stranded gas that is also associated gas).

Multiple opportunities exist to achieve better value for natural gas than its local fuel value. First, the gas or its gaseous components can be exported to markets having higher fuel values. Exportation can be performed by pipeline, as pressurized gas, or by ocean tanker as liquefied natural gas (LNG, comprised of 99% methane) or natural gas liquids (NGL, comprised of C₂–C₄ alkanes). Second, GTL technologies can be used to chemically convert natural gas to higher value liquid fuels such as naphtha or diesel, using F-T technologies; or to fuel grade methanol, DME (directly or via methanol), MTBE (via methanol), or gasoline (via methanol); or to chemicals.

A flowchart of conceptual routes for chemical conversion of the methane in natural gas to liquids is shown in Fig. 1. All current gas to liquids chemical production processes require that natural gas is first converted to synthesis gas (syngas), a combination of hydrogen and carbon monoxide. The syngas is subsequently converted in a synthesis section to product. GTL processes of today are therefore of the indirect type, in that they proceed first through the intermediate syngas instead of converting methane directly to liquid product. Direct syntheses of fuels from natural gas include oxidative coupling of methane and selective oxidation of natural gas to methanol. These have been investigated in the past but poor economics to date have prevented their commercialization.

The intermediacy of synthesis gas is common when converting various hydrocarbonaceous feedstocks such as natural gas, naphtha, residual oil, petroleum coke, coal, or biomass. The lowest cost routes for syngas production are based on natural gas. There are several different approaches to converting natural gas to syngas, the primary methods being catalytic steam methane reforming (SMR), autothermal reforming (ATR), gas-heated reforming (GHR), partial oxidation (POx), heat exchange reforming, and variants thereof. The ratio of H₂/CO produced varies with the process design as well as the H/C ratio in the feedstock. A process will tend to be most economic if the H₂/CO ratio produced is well matched to the ratio needed in the next step. Exceptions include processes configured to produce more H₂ than needed, because this extra H₂ can be separated and used for power or to augment the syngas



Gas to Liquid Technologies. Figure 1

Conceptual routes for chemical conversion of gas to liquids [7]

ratio of another chemical process. If the H_2/CO ratio is lower than needed, then an additional source of the expensive component H_2 must be supplied. With an H/C ratio of 4, methane is well suited for conversion to syngas when the final product slate will contain saturated hydrocarbons or methanol.

The synthesis of hydrocarbons from CO hydrogenation over transition metals was discovered in 1902 when Sabatier and Sanderens produced CH_4 from H_2 and CO mixtures passed over Ni , Fe , and Co catalysts. In 1923, Fischer and Tropsch reported the use of alkalized Fe catalysts to produce liquid hydrocarbons rich in oxygenated compounds – termed the synthol process. Succeeding these initial discoveries, considerable effort has gone into developing catalysts for the process to produce liquid hydrocarbons. Eventually, a precipitated Co catalyst promoted with ThO_2 and MgO supported by Kieselguhr (diatomaceous earth) became known as the standard atmospheric process catalyst. In 1936, Fischer and Pilcher developed the medium pressure (10–15 atm) Fischer–Tropsch synthesis process. Following this development, alkalized Fe catalysts were implemented into this medium

pressure process. Collectively the process of converting CO and H_2 mixtures to liquid hydrocarbons over a transition metal catalyst became known as the Fischer–Tropsch synthesis.

Methanol is a commodity chemical, and one of the top ten chemicals produced globally. The long time interest in methanol is due to its potential use as a fuel and as a feedstock to the chemicals industry. In particular, methanol can be used directly or blended with various petroleum products as a clean burning transportation fuel. Methanol is also an important chemical intermediate used to produce formaldehyde, dimethyl ether, methyl tertiary-butyl ether (MTBE), acetic acid, methylamines, and methyl halides among others.

Production of methanol began in the 1800s, with the isolation of wood alcohol from the dry distillation (pyrolysis) of wood. Research and development efforts at the beginning of the twentieth century involving the reaction of syngas to give liquid fuels and chemicals led to the discovery of a methanol synthesis process (concurrently with the development of the Fischer–Tropsch synthesis). Methanol synthesis

is now a well-developed commercial catalytic process with high reaction rate and selectivity (up to 99%). For economic reasons, methanol is produced almost exclusively (over 90%) via the reforming of natural gas.

Survey of Specific Gas to Liquids Technologies

This survey of GTL processes covers the subtopics of liquefied natural gas, natural gas liquids, methane reforming, Fischer–Tropsch GTL, and methanol synthesis from natural gas.

Liquefied Natural Gas (LNG)

Liquefaction of natural gas reduces its volume to about 1/600 the volume of the gas as measured under standard conditions. The most compelling justification for the production of LNG is this volume shrinkage that makes LNG, unlike natural gas itself, practical to both store and transport.

Production of LNG has been commercially practiced since 1960. The two major applications of LNG technology are:

- (a) Storage of LNG for use in peak shaving plants with seasonal adjustment (mostly in the USA)
- (b) Base load LNG plants for international trade with shipping from remote areas to developed countries via dedicated LNG ocean tankers

The capacities of peak shaving plants are more than an order of magnitude smaller than base load LNG plants.

Construction of pipelines depends on the relative values for gas in the supplying and consuming regions. Plans exist for pipeline delivery of gas from isolated eastern Russian gas fields to consuming markets in Asia (Japan, Korea, and China). Nonetheless shipping tends to be preferred over long-range pipeline construction, because pipeline costs are highly capital intensive: a 621 mile (1,000 km) long pipeline would cost some \$1 billion, depending on ground conditions. Exporting LNG to Asia and Europe from distant production fields has become economic as a result of improvements to thermodynamic efficiencies of LNG facilities. However, shipping is still expensive at a cost of at least \$15 per bbl to transport from the reservoir to the consumer's storage tanks.

Recent industry trends indicate that tankers to hold as much as 135,000 m³ (equivalent to 3.2 billion scf of natural gas) have been specially built, with built capacities expected to soon increase to 165,000 m³. For a shipping distance of 4,000 or 6,600 miles, the shipping requirements for a 1 billion scf/day liquefaction plant can be met by fleets of six or nine tankers, each tanker making 17 or 11 trips per year, respectively. Following transport of the LNG by tanker from the liquefaction site, it is off-loaded at a shore terminal. Vapor generated is compressed for injection into the local distribution pipeline. Independently, a continuous regasification of the LNG from storage tanks is performed, and the amount used for the time period between ships must obviously match the amount unloaded. The LNG pressure is raised by pump to the pressure necessary for it to be vaporized into the distribution pipeline.

Plant Considerations

The liquefaction energy required in a LNG plant typically has been reported as 9–12% of the heat energy in the natural gas, and 9–10% energy shrinkage is a typical number for the modern mega-tonnage capacity plants (without combined cycle systems). Because LNG is stored and delivered at atmospheric pressure and –160°C (–256°F), compression and very deep refrigeration are needed, with associated large consumption of energy. LNG projects have a very high capital cost, in the range of \$1.0–1.5 billion for a 3–3.3 million tons per year train on a greenfield site, with 45–60% attributed to off-sites and infrastructure depending particularly on requirements for LNG storage, the marine system, and the heat rejection method. The technology is relatively mature, although economies of scale are being investigated to increase scale size to 4.5–5.5 million tons per year.

The thermal efficiency of LNG plant is determined by two major factors:

- (a) Refrigeration cycle efficiency
- (b) Power cycle efficiency

Increasing the thermal efficiency of an LNG plant for a given turbine/driver configuration will decrease the cost of production and also minimize both site gas consumption and greenhouse gas emission of CO₂

generated by combustion (0.20 t CO₂/t of LNG at high thermal efficiency, versus 0.25–0.35 t CO₂/t in a typical LNG plant) [8] as well as providing low NO_x emission (0.095 kg/t).

The key capital cost elements in LNG facilities are, in descending order:

1. Gas turbines, steam turbines, or motor drivers for refrigeration service
2. Refrigeration compressors, typically over 100,000 kW
3. Steam and power generation including turbines waste heat recovery
4. LNG storage, typically over 250,000 m³ capacity per tank (equivalent to approximately 112,500 t LNG per tank)
5. LNG loading terminal including jetty or causeway
6. Heat rejection system, in most cases, as suggested, seawater cooling
7. Cold box and prechilling for gas liquefaction
8. Natural gas pretreating for CO₂ removal gas drying and mercury adsorption
9. LPG and natural gasoline recovery as by-products
10. Fuel gas cold recovery and compression

Although heat transfer for chilling (i.e., the cold box) is less important than the capital investment associated with compression, the correct selection of the optimized refrigeration cycle and associated drivers affects both the compression and heat rejection systems and economics.

The most common refrigeration system in LNG plants involves prechilling with propane followed by use of the mixed refrigerant system. Close to 90% of these plants are licensed by Air Products Corporation, Inc. (APCI). The mixed refrigerant liquefaction systems use a mixture of mostly methane and ethane, in about 1.2–2.0:1 molar ratio. Depending on the feed gas composition, up to 3 mol% nitrogen and 6–12 mol% propane may be added to optimize the LNG boiling curve. The refrigerant cooling curve is adjusted to follow closely the feed gas cooling curve in order to achieve maximum thermodynamic efficiency [9].

One significant exception to the propane prechilling, mixed refrigerant approach is the Phillips Kenai Peninsula plant in Alaska, which commenced operation in 1969. This plant produces about 1.5

million tons per year of LNG by a cascade refrigeration system using pure propane, ethylene, and methane refrigeration cycles. The efficiency of this cascade system has been reported as at the lower end of the scale, about 88% [10]. It is reasonable to assume that the adiabatic efficiency of the refrigeration compressors currently operated by ConocoPhillips in Kenai are on the order of 70%, versus adiabatic efficiencies of 80–85% of more modern centrifugal compressors with three-dimensional blades. Even so, the plant has proved reliable and profitable.

Several patents issued to Phillips (now ConocoPhillips) suggest that improvements in cycle efficiency would result from improved refrigeration load distribution, nitrogen stripping (for nitrogen rich gas), open loop methane refrigeration, and LPG recovery [10–12]. The technology appears driven by the desire to distribute load equally among the propane cycle, ethylene cycle, and methane cycle. This is achieved in part by superheating the methane and ethylene refrigerant vapors fed to the compressors, probably to about –46°C (–50°F). Design and construction considerations then allow the use of six identical gas turbines, such as frame 5D (nominal 30,000 kW), to be used. The open loop methane refrigeration appears also to use the methane refrigeration compressor as the fuel gas compressor.

The loads of the propylene compressor are about 47.3%, the refrigeration ethylene compression loads are about 36.4%, and the methane compressor loads are 16.3% of the total refrigeration load. In practice the combined load on the steam turbines may be about 27% of the total motive power in the facility. The estimated power consumption of the Phillips system, for feed gas at 38°C (100°F) and 650 psig, is 371 kWh/t, assuming very lean gas, such as 96 vol% methane, at a heat rejection temperature speculated to be 38°C (100°F) [10]. This estimate probably includes 15 kWh/t of fuel gas compression [13].

A potential drawback of load equalization is use of up to 5% additional refrigeration power and, in case of fixed speed gas turbines, a reduction in production capacity [13]. Further, this could result in a more complex refrigeration cycle where the propane cycle, ethylene cycle, and methane cycles must be more highly

heat integrated [10]. Despite potential drawbacks in power consumption and capacity reduction, the concept of cascade refrigeration may be very sound [14], given the following considerations:

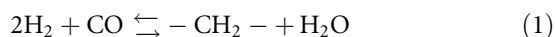
- Even the older, closed loop methane refrigeration as used in Kenai could have an advantage by allowing higher suction pressure to the methane compressor, about 1.7 kg/cm² (24 psia) instead of an estimated 1.1 kg/cm² (16 psia) for an assumed open loop compressor.
- Modern design methodology would allow the use of conventional carbon steel metallurgy in the compressors.
- Just one electric motor-driven fuel gas compressor is used for two trains.
- In case of outage of the fuel gas compressor, backup is provided by a draw from the feed gas.
- The ethylene refrigeration cycle, aside from superheating the suction to the first stage to about −46°C (−50°F), apparently comprises only a single side load with a probable goal of obtaining the compression in a single casing – potentially without employing superheating of the suction to the ethylene compressors.
- Use of two side loads reduces refrigeration load by 2% and increases production capacity by 2–3% (although using two casings also increases capital investment, possibly by about \$5 million).
- On average about 6.2% of the total power is exported as electric power outside the boundary limits. (During times of high ambient temperature, the start-up steam turbine attached to the gas turbine increases its relative steam consumption and electric power export drops to near zero).

Given the pros and cons of cascade versus conventional mixed refrigerant systems, an objective comparative evaluation could be made only on a case-by-case, site-specific basis for LNG production.

Methane Reforming in F-T GTL

Reforming is the means by which natural gas is converted to the synthesis gas used as feed for GTL processes. Syngas for use in Fischer–Tropsch GTL is best characterized by the H₂/CO ratio, which should

be about 2.0, after reforming, as per the generic F-T stoichiometry of Eq. 1.



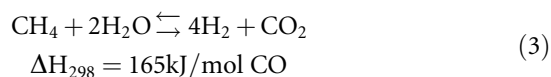
There are many process options for reforming technology, used alone or in hybrid reforming approaches, and these are compared in Table 1.

The choice of how syngas is produced depends on several factors, including:

- Match between the H₂/CO ratio produced and the H₂ demand of the overall process
- Plant and process scale
- The need for, or feasibility of, using an air separation unit (ASU) or oxygen enrichment facilities in the overall process
- Scale and logistics of capital equipment, such as compressors and other gas-handling equipment
- Heat integration and gas recycle options

Steam Methane Reforming (SMR)

Steam methane reforming, or steam reforming for short, is a catalytic conversion of natural gas by reaction with steam. The basic chemistry of steam methane reforming is shown in Eqs. 2 and 3.



These reactions are quite endothermic, requiring input of significant transferred heat, and SMR is therefore carried out at high temperature with the catalyst inside tubes within a fired furnace. The amount of steam used is in excess of the reaction stoichiometry requirements, as required to prevent the catalyst from coking.

SMR is the dominant reforming technology used for the production of methanol, ammonia, and other petrochemical products produced from methane at large scale. One reason is that the SMR process is scalable. Another reason is that the high H₂/CO ratio produced in SMR is stoichiometrically suitable for downstream formation of saturated hydrocarbons

Gas to Liquid Technologies. Table 1 Comparison of technologies for syngas generation from natural gas [15]

Technology	Advantages	Disadvantages
Steam methane reforming (SMR)	<ul style="list-style-type: none"> • Most extensive industrial experience • O₂ not required • Lowest process temperature • (Best H₂/CO ratio if producing H₂) 	<ul style="list-style-type: none"> • H₂/CO ratio higher than typically required for syngas production • Highest air emissions
Gas-heated reforming (GHR), heat exchange reforming (HER)	<ul style="list-style-type: none"> • Compact size and footprint • Application flexibility offers options for incremental capacity 	<ul style="list-style-type: none"> • Limited commercial experience • Usually best coupled with another syngas generation technology, such as ATR
Two-step reforming (SMR followed by O ₂ -blown secondary reforming)	<ul style="list-style-type: none"> • Size of SMR is reduced • Low methane breakthrough favors high purity syngas • Methane content of syngas can be tailored by adjusting secondary reformer outlet temperature 	<ul style="list-style-type: none"> • Increased process complexity • Higher process temperature than SMR • Usually requires O₂
Autothermal reforming (ATR)	<ul style="list-style-type: none"> • Typical H₂/CO ratio produced is close to stoichiometric for F-T and methanol syntheses • Lower process temperature requirement than Pox • Low methane breakthrough • Methane content of syngas can be tailored by adjusting reformer outlet temperature 	<ul style="list-style-type: none"> • Higher process temperature than SMR • Usually requires O₂
Partial oxidation (Pox)	<ul style="list-style-type: none"> • Feedstock desulfurization not required • Absence of catalyst permits carbon formation and, therefore operation without steam, significantly lowering syngas CO₂ content • Low methane breakthrough • Low H₂/CO ratio advantageous where ratio < 2.0 is required, such as dimethylether synthesis 	<ul style="list-style-type: none"> • Low H₂/CO ratio disadvantageous where ratio > 2.0 is required • Very high process operating temperatures • Usually requires O₂ • High-temperature heat recovery, and soot formation and handling, add process complexity • Low methane content of syngas not easily modified to meet downstream processing requirements

and oxygenates. The theoretical ratio for H₂/CO of 3:1 is not reached in practice, the maximum practical value being about 2.8, but in the absence of complex recycle schemes, SMR provides the highest H₂/CO ratio available in reforming of natural gas.

Conventional steam reforming catalysts are 10–33 wt% NiO on a mineral support (alumina, cement, or magnesia). Although natural gas usually contains only small amounts of sulfur compounds,

generally in the form of H₂S, sulfur compounds are the main poisons of reforming catalysts. Use of uranium oxide or chromium oxide as promoters can impart higher tolerance to sulfur poisoning, but even at a sulfur concentration of 0.1 ppm the catalyst can begin to deactivate and the best practice is to remove sulfur from the raw feed. To maintain a 3-year catalyst lifetime, the sulfur concentration in the reformer feed gas should be less than 0.5 ppm. If the sulfur

concentration in the raw feed gas is greater than 1%, the sulfur must be removed by chemical or physical scrubbing. An upfront desulfurization unit is used to absorb H_2S onto a ZnO bed. Any remaining organic sulfur compounds and carbonyl sulfide are partially cracked and absorbed on the zinc oxide bed.

Recent improvements to catalysts and the reforming process have allowed design for operation of side fired reformers in SMR, at conditions not possible with other reforming methods. The most notable developments are noted below.

- Introduction of new generations of catalysts, suitable for pre-reforming and for reforming of heavy feedstocks

Pre-reforming catalysis converts higher hydrocarbons (C_{2+}) in the natural gas feed into a mixture of methane, hydrogen, and carbon oxides.

- Commercial application of advanced reforming at low steam/carbon (S/C) ratios down to 1.5, and high outlet temperatures above 950°C

Low S/C ratios improve energy efficiency.

- Significant increase of heat flux

Operation at average heat flux above $150,000 \text{ kcal/m}^3 \text{ h}$ has been demonstrated in Haldor Topsøe's Houston, Texas, process development unit (PDU). A high flux unit is now in commercial operation.

- Increase of unit capacity

Steam reformers can now be constructed for single train reforming units with capacity to provide syngas for a methanol plant of up to 3,000 t/day production.

- Heat exchange reforming

Heat exchange reforming has been developed and commercialized including with gas-heated option, where the steam reforming reaction is carried out in a unit heated by the product gas from an autothermal reformer.

An obvious advantage of SMR is that it does not need an oxygen plant. However, since steam reformers are more costly than either POx or autothermal reformers, there is a minimal plant size above which the economy of scale of a cryogenic oxygen plant in combination with a POx or autothermal reformer is less expensive than a steam reformer on its own. Use of SMR for Fischer–Tropsch applications does involve

recycling of CO_2 and removal of excess hydrogen, by means of pressure swing absorbers or membrane separators, to lower the H_2/CO ratio to a level acceptable to the Fischer–Tropsch stoichiometry. Due to the costs involved with these steps, it is most likely that steam reforming will only be considered for F-T GTL when one or more of the following conditions hold:

- A relatively small GTL plant, with a capacity well below 10,000 bbl/day.
- The additional hydrogen can be used for other applications like methanol or ammonia production.
- The natural gas feed has a high CO_2 content.
- A suitable water supply can be obtained at low cost.

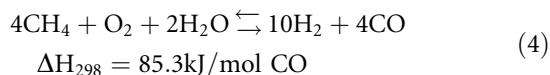
Compact Reforming

Compact reforming represents a novel mechanical design approach to conventional SMR. Under development by BP, it is currently undergoing extended test runs at their new 300 bbl/day GTL demonstration facility in Alaska. The reformer design resembles a conventional shell and tube heat exchanger, and SMR reactions occur tube side of this reactor, with the tubes filled with conventional Ni-based catalyst. Heat for the endothermic SMR reactions is provided shell side, where the tubes are heated by combustion of a fuel and air mixture. Heat transfer is claimed to occur more efficiently in what is described as a highly countercurrent device.

The shell side combustion zone may also be under elevated pressure, leading to more efficient convective heat transfer to the tubes, and this plus a considerably lower catalyst quantity required considerably reduces the size relative to the conventional SMR furnace design configuration – hence the term “compact” reforming coined by BP. The principal advantages claimed for this approach compared to a conventional SMR are a major reduction of capital cost along with an improvement in thermal efficiency. The reduced space requirements are claimed to make this reforming technology more amenable to installation on barges or oceangoing “plantships” for the conversion of relatively small or temporary sources of natural gas. A drawback of the current state of compact reforming technology is that a large number of parallel units would be required for a world-scale GTL plant.

Autothermal Reforming (ATR)

Autothermal reforming (ATR) has been recognized since the mid-1990s as the technology of choice for production of synthesis gas at large scale. Conceptually ATR is similar to catalytic SMR except that ATR includes the use of oxygen, which renders the reaction less endothermic on a per mole basis than SMR. The stoichiometry for ATR is shown in Eq. 4.



In practice the product distribution is strongly affected by the choice of steam/carbon (S/C) ratio. At a S/C ratio of 1.3, the syngas will have a H_2/CO ratio of about 2.5, which is higher than the ratio of about 2.0 needed for Fischer–Tropsch synthesis. The H_2/CO ratio can be controlled by a combination of lowering the S/C ratio and adding CO_2 or CO_2 -rich gas to the ATR feed, such as by recycling CO_2 to the reformer. Decreasing the S/C ratio in the feed gas decreases the amount of CO_2 required when adjusting the synthesis gas composition to the desired ratio for F-T synthesis.

A typical ATR process concept for the production of synthesis gas includes the fundamental steps of desulfurization, adiabatic pre-reforming, ATR, and heat recovery. In the desulfurization section, the sulfur present in natural gas feedstock is removed to avoid poisoning the downstream pre-reforming catalyst. Steam, and optionally recycled CO_2 , is added to the desulfurized natural gas and, after further heating, the resultant mixture is passed to a pre-reformer. In pre-reforming, preheating of the gas is done in a fired heater upstream of the ATR. The use of a pre-reformer upstream of the ATR unit reduces the oxygen consumption per unit of syngas produced during ATR. The exit gas from the ATR is cooled by high-pressure steam production and boiler feed water preheat, and CO_2 is removed as needed, using methyldiethanolamine (MDEA, N-methyl-diethanolamine) as an absorbent in a CO_2 capture unit. After CO_2 removal, the synthesis gas with a H_2/CO ratio of 2.0 is available for Fischer–Tropsch synthesis. Various process variants are possible, for example, adjustment of synthesis gas composition by recycling CO_2 -rich tail gas from the Fischer–Tropsch synthesis.

The result is that ATR offers a reduction in the investment per barrel of product, and the possibility of higher single-line capacity compared to conventional steam reforming. Optimized design in combination with reduction of the S/C ratio can be expected to result in a major increase in the single-line capacity, by more than 25% within the next few years. Currently single train methanol plant units based on ATR can be designed with capacities up to about 7,000 t/day.

The key component of the ATR process is the ATR reactor itself. The ATR reactor has a compact design consisting of a burner, combustion chamber, and a catalyst bed placed in a refractory lined pressure vessel. The pre-reformed natural gas reacts with oxygen and steam in a sub-stoichiometric flame. In the catalyst bed, the gas is equilibrated with respect to the methane steam reforming and shift reactions. Product gas composition is determined by the thermodynamic equilibrium of these reactions at the exit temperature and pressure of the reactor. The exit temperature is determined by the adiabatic heat balance based on the composition and flow of the feed, steam, and oxygen added to the reactor. The product gas is completely free of soot and oxygen.

It is essential that the combined design of burner, catalyst, and reactor ensures that the precursors are destroyed by the catalyst bed to avoid soot formation. The soot limit, that is, the S/C ratio at which soot formation starts, has been investigated by experiment at many combinations of temperature, pressure, so that prediction of acceptable operating conditions for specific feedstocks is now possible.

Additional design parameters of the autothermal syngas generation system that influence cost and thermal efficiency include:

- The preheat temperatures of oxygen and natural gas
The higher these temperatures are, the less oxygen will be used. The maximum preheat temperatures are determined by safety factors and by the need to prevent soot formation.
- The pressure of the steam generated in the waste heat reboiler

The higher the steam pressure, the more efficiently can energy be recovered from steam, but the more costly the steam and boiler feed water treatment systems become. The optimum steam pressure will be determined by the relative cost of capital and energy.

The feasibility of soot-free operation of the ATR reactor at very low S/C ratio was demonstrated in Haldor Topsøe's Houston development unit in 1997–1999, with further testing performed since then to explore the limits of the ATR technology with respect to feed gas composition, pressure, and temperatures. Operation at industrial scale at a S/C ratio of 0.6 was demonstrated in a South African plant in 1999 and has been in operation in Europe since the start of 2002 [16]. Two lines have been in operation at Sasolburg since 2004. Haldor Topsøe's ATR technology was also selected for the Oryx plant that Qatar Petroleum and Sasol began operating in 2006 and includes a 6,000 mtd syngas generator; for Chevron Nigeria's Escravos, Nigeria GTL plant scheduled for completion in 2011; and for the methanol portion of Eurochem's gas-to petrochemical complex planned in Lagos State, Nigeria.

Sasol, Syntroleum, and ExxonMobil each have their own proprietary versions of ATR. The most prevalent layout of the syngas production section is a combination of adiabatic pre-reforming and fixed bed ATR. This layout results in high flexibility for variations in natural gas feed and, if needed, in tail gas recycle gas compositions, as well as the ability to reduce oxygen consumption per unit of syngas produced.

ExxonMobil employs a fluidized bed catalytic reformer. Steam-diluted oxygen is fed to the reactor through nozzles, separately from the natural gas. The oxygen reacts exothermically with natural gas in a burning zone near the oxygen inlet. The balance of steam needed in the process is admixed with natural gas and fed at a level below the oxygen inlet to the autothermal reaction zone. A mixture of nickel on alumina catalyst may be combined with a solids diluent (alumina) to form the fluidized bed. The fluidized bed reactor provides superior characteristics of heat and mass transfer for autothermal reforming.

Another ATR approach under consideration is to use slurry bed F-T reactors in series within each train, as the slurry system would facilitate the application of ATR technology at large scale (>10,000 bpd). With this approach, water is removed between stages to enhance the reactant partial pressures in the downstream reactor stage. The removal of other diluents (such as carbon dioxide) between stages may also be considered. This allows lower recycle ratios, increased

steam production from the reactor heat removal system, and decreased overall reactor volumes. These advantages will not necessarily compensate for the increased complexity of needing to use reactors in series at larger scale, and detailed studies are required to determine the most cost-effective design approach.

Currently, the most attractive ATR syngas generation technology appears to be oxygen-blown ATR at low S/C ratio. This technology has been chosen for F-T GTL projects that are closest to realization. Extensive testing has proved that technology is ready for application at a large industrial scale with S/C ratios down to 0.6. Application of even lower S/C ratios has been demonstrated at small scale, and the next generation ATR scheme is under development to cut the S/C ratio even lower, to 0.4 or possibly 0.2 [17]. Such low S/C could decrease oxygen requirements and boost the single-line capacity still further.

Two-step reforming is a variant of ATR wherein the exit gas from a fired tubular reformer is further processed in an oxygen-blown, secondary reformer. In both steps, the ATR is carried out in a refractory lined vessel containing a mixer/burner and a catalyst bed. Conditions leading to soot formation are rigorously avoided.

Gas-Heated Reforming/Heat Exchange Reforming (GHR/HER)

Another technology that produces suitable synthesis gas for GTL units is a combination of steam reforming and ATR, performed in separate units, in a process referred to as gas-heated reforming, GHR, or heat exchange reforming, HER. The concept is to use heat available from process gas downstream of the ATR for steam reforming, in a heat exchanger type reactor. The purpose is to reduce oxygen consumption and increase carbon efficiency by optimizing the recycle of tail gas – beyond the range possible with ATR alone – and to achieve this without additional firing that would create an undesired excess of water. The gas-heated reformer is a compact alternative to the fired steam reformer wherein high-grade heat is recycled while the S/C ratio remains below 0.6 [18].

Conventional reformers provide the needed heat by combustion of fuel at atmospheric pressure in a refractory lined duct. Because the convective heat

transfer coefficient of a GHR is considerably higher at the elevated pressures of its hot gas side, unit size reductions claimed can be 10–15 times smaller than conventional tubular steam methane reformers. The absence of combustion flue gases results in an inherently higher thermal efficiency for a GHR, since no combustion heat is lost with the warm flue gases discharged to the atmosphere from the stack. Furthermore, the reduced power consumption, which results from eliminating the need for the associated steam methane reformer air and flue gas blowers, contributes to the overall energy saving of a GHR. Much of this thermal efficiency advantage may be offset, however, by the need to generate additional steam externally in an off-site boiler in order to compensate for the lost waste heat recovery potential of a conventional reformer system.

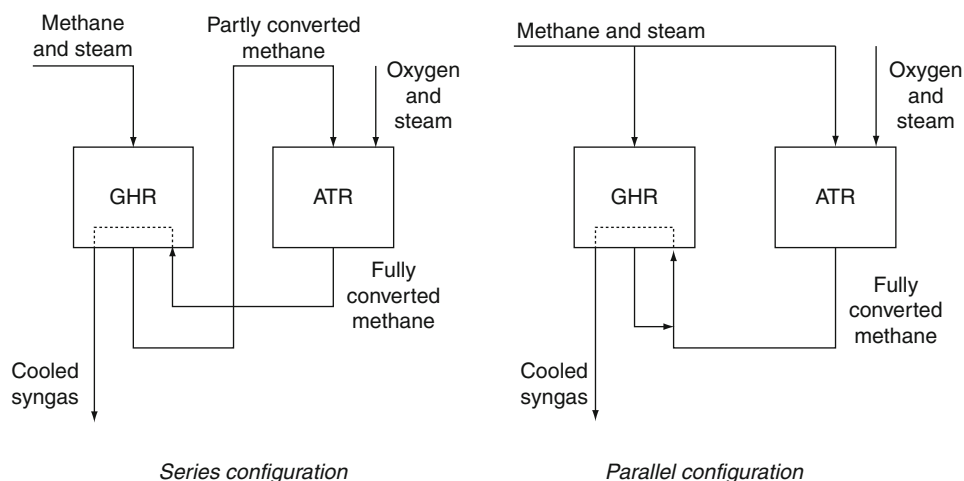
GHR can be combined with ATR to reduce energy consumption and enable the use of smaller air separation plants, while reducing CO₂ emissions possibly by as much as 50%. In addition, these GHR/ATR schemes allow shrinking of the required high-pressure steam system and reduction in size of steam turbines. Both series and parallel arrangements of GHR with ATR have been studied and are depicted in Fig. 2 [18]. In the series arrangement of HER/GHR, all gas passes through the steam reforming unit and then the ATR. The lower limit for the S/C ratio is typically

determined by the steam reforming catalyst [19]. The parallel arrangement allows individual optimization of the S/C ratio by feeding the two reformers independently.

GHR has been employed since the 1980s in some ammonia and methanol plants. In the past, highly efficient GHR was not viewed as an option for high capacity syngas production, mainly because of concerns about metal dusting corrosion. This reformer must operate at higher temperatures in the parallel arrangement than in the series arrangement, to obtain a low methane leakage. Higher operating temperatures may lead to higher cost and more severe conditions, and the main challenge in both series and parallel arrangements is preventing metal dusting corrosion on the shell side of the heat exchange reformer. Mechanical design, including choice of materials of construction, is critical.

A heat exchange reformer designed by Haldor Topsøe for operating in tandem with an ATR at a low S/C ratio was started up at Sasol's facilities in Secunda, South Africa, in early 2003 [16, 20]. The project at Secunda (also referred to as "process gas-heated reforming" or PGR) involved revamping an existing ATR for an increase in syngas production capacity of 30% via GHR.

Engineers at Johnson Matthey have stated that with mechanical design and configuration improvements,



Gas to Liquid Technologies. Figure 2

Series and parallel configuration of the combination of GHR and ATR [18]

GHR/ATR technology is likely to deliver better efficiency and lower operating costs than conventional ATR for F-T GTL plants of the future. In addition, Air Products and Toyo Engineering Corporation have each studied GHR technology. Although commercial F-T GTL plants using ATR/GHR combinations have yet been built, retrofit projects at large-scale methanol and GTL plants are possible, as well as potential applications of ATR/GHR in refinery hydrogen plants.

Air Blown Autothermal Reforming

It is possible to use air instead of purified oxygen as oxidant in autothermal reforming, and an F-T synthesis process based on air blown ATR was developed by Syntroleum Corporation and demonstrated in their plant in Tulsa, Oklahoma, in 2002–2003. In a very similar operation, air blown secondary reforming is used extensively in the production of synthesis gas in ammonia manufacture. The main advantage for air blown processes is that no air separation unit will be required, which could be a strategic advantage for smaller scale, offshore, barge mounted F-T GTL units. For large-scale land-based units, it may present certain restrictions on the energy system configuration of the plant. The savings of eliminating the capital cost of an air separation cost will be offset by the need for compression of incoming air, for example, which will result perhaps in as much as 50% more power required than that required to produce oxygen.

In addition, the increased volumetric flow of nitrogen-diluted syngas will make all downstream equipment through the F-T synthesis reactors larger. Also, the high content of nitrogen in the synthesis gas excludes the use of internal recycle in the F-T synthesis section and external recycle of tail gas from the F-T reactor back to the syngas section of the plant is excluded. This means that all tail gas must be used as fuel, which because of the high nitrogen content will be of poor quality. At a conversion of hydrogen and CO in the synthesis gas of about 70%, the heating value of the tail gas will be so low that it will not burn unassisted. Either supplemental fuel or catalytic combustion will be required if the energy content in the tail gas is converted to power, and a large amount of power will

be generated for export. This may be desirable, however, in remote locations where power may be in short supply.

Partial Oxidation (POx) and Catalytic Partial Oxidation (CPOx) Reforming

When the quantity of oxidant (O_2) added to methane is very low and the reaction temperature is raised significantly, uncatalyzed partial combustion of pretreated natural gas occurs. This homogeneous reaction is referred to as gasification or partial oxidation (POx). The resulting syngas is quenched or cooled by steam production, and carbonaceous by-products such as soot are removed by washing. The formation of carbonaceous by-products influences carbon efficiency. For partial oxidation, the syngas H_2/CO ratio produced is slightly below 2.0, which is closest to the optimum needed for the Fischer–Tropsch or methanol syntheses. This low H_2/CO ratio gas results from very little, if any, steam use in the process. Due to the absence of catalyst, the reformer operates at an exit temperature of about 1,300–1,400°C. This high temperature and the absence of catalyst have the following disadvantages as compared to an autothermal reformer.

- Formation of soot and much higher levels of ammonia and HCN appear in the syngas, compared to ATR, which necessitates the use of a scrubber to clean the gas.
- Higher consumption of oxygen.
- Due to the absence of a water gas shift reaction with the POx catalyst, the unconverted methane as well as the methane produced by the Fischer–Tropsch reaction cannot be recycled to the reformer without first removing the CO_2 from the Fischer–Tropsch tail gas.

Depending on the energy needs of the plant, the syngas from the POx reformer can either be cooled by means of a water quench or by the production of steam in a heat exchanger. A quench system is the less costly of the two approaches but is also less thermally efficient. In designing a POx-based GTL plant, the choice between a quench or a waste heat boiler will depend on the relative cost of capital and energy. As mentioned above, CO_2 -rich tail gas separated from the syngas may be recycled to the partial oxidation reactor to improve

carbon yield. If recycle is used, the gas production in the POx units must be supplemented by high hydrogen content syngas in an auxiliary steam reformer, as in the Shell process. In Shell's Bintulu GTL plant, each POx unit originally had a capacity rating for production of 3,000 bbl/day of F-T products. Shell has recently scaled up their gasification units to enable production of 10,000 bbl/day of F-T product equivalents.

When a catalyst is employed to assist the partial oxidation, the process is referred to as catalytic partial oxidation (CPOx). The catalyst increases the activity of natural gas conversion and, therefore, provides options for higher throughput, lower operating temperature, or both.

CPOx with no burner (flame) was first practiced on a small scale at low pressures in the 1950s by Haldor Topsøe, and then later with higher pressures by Lurgi. Early versions of CPOx used premixing of hydrocarbon/steam feed and oxygen, with an ignition catalyst placed on top of a fixed bed catalyst, similar to the catalyst used in an ATR-based unit and operating with similar space velocities. Today's CPOx versions feature a combined catalyst system performing both ignition and partial oxidation. The noble metal catalyst operates at very low residence time (very high space velocity). The hydrocarbon feed stream, oxygen, and steam are mixed upstream from the catalytic reactor. All chemical conversion takes place in the reactor and no burner is used. The exit gas composition from CPOx is similar to the exit gas from ATR, given like amount and properties (temperature, pressure, composition) of the inlet streams.

The main potential advantage of catalytic partial oxidation is the smaller size of the reactor. There are also potential disadvantages that will likely outweigh the savings in the cost of the reactor, which is rather modest relative to the total investment in the syngas generation section. The main disadvantages of CPOx relate to the premixing of natural gas hydrocarbon feed and oxygen. The need to premix excludes combinations of CPOx linked in a series arrangement, in particular with pre-reforming or heat exchange reforming, for safety reasons. Premixing hydrocarbon feed with oxygen limits the applicable reactor preheat temperatures to rather low values, kept below the self-ignition

temperature of the mixture to maintain control when premixing oxygen and hydrocarbon. A feed gas mixture containing methane and oxygen may spontaneously ignite at temperatures above the autoignition temperature, depending on the actual gas properties, and the industry does not yet have confidence that this issue has been adequately addressed.

When consumption of natural gas and oxygen are compared for CPOx and a combination of pre-reforming and ATR, consumption of both gases are significantly higher for CPOx than for the combination of pre-reforming and ATR. The main reason is that with CPOx operating at a lower inlet temperature than ATR (250–450°C vs. 600°C), a significant part of the hydrocarbon feed is combusted to CO₂ and steam (using the larger amount of oxygen) to generate heat to reach the desired syngas exit temperature of 1,050°C.

Considering that a large portion of the syngas process section of a F-T plant is for the air separation unit, CPOx does not appear to be economical under these conditions. CPOx was being investigated by ConocoPhillips but they have recently discontinued this approach.

An alternative for CPOx could be the use of air as an oxidant. This alternative is being pursued by many as a basis for fuel processing systems for fuel cells. For GTL and similar applications operating at high pressures, however, the issue of safety still remains even with air, although to a lesser extent than with oxygen. Oxygen consumption may be conserved (to an amount about 10% less than the amount required for the ATR reactors) if the inlet temperature can be safely raised to a higher level (about 480°C) with a lower oxygen content in the hydrocarbon/oxygen mix. Additional operational issues with CPOx include concerns about catalyst life, carbon dusting, and combustion "flash back" potential in an integrated F-T GTL process plant.

The combination of POx and conventional steam reforming appears to be the nearest competitor to ATR technology. However, catalytic partial oxidation, either oxygen or air blown, does not appear to offer sufficient advantages to outweigh its limitations with respect to operating conditions and inherent safety concerns. CPOx seems to require an oxygen consumption boost, and it is unclear whether this need can be

adequately diminished. CPOx is not expected to be important in the GTL industry in general, even if successfully developed for commercialization.

Catalytic Membrane Reactors

Catalytic membrane reactors (CMR) are based on the concept that selective transport of reactant(s) or product(s) can influence an equilibrium-limited reaction to proceed further in the forward direction. The combination of SMR and the WGS is equilibrium limited, and H₂-selective membranes have been used to lower the temperature required for SMR by facilitating the selective removal of H₂ as it forms. Both dense (mostly palladium) and porous (such as silica or alumina) H₂-perselective membranes have been studied [21]. Throughput limitations, membrane stability, and cost combined with the ready availability of several other reforming technologies have thus far hampered commercial development of H₂-permselective membrane reactors for reforming applications.

Another approach to CMR for methane reforming allows a selective passage of reactant oxygen. Nonporous membranes prepared from perovskites or similar ceramic materials allow ionic transport of oxygen from air at low pressure, with up to 100% selectivity for O₂. The selectively transported oxygen moves into a reaction mixture at low partial pressure of oxygen, against a large difference in total pressure.

Process concepts based on the use of O₂-selective CMR eliminate the need for an air separation unit (ASU) and associated large compressors. Successful elimination of the ASU would impart significant economic potential for this technology; however, difficulties with CMR would need to be overcome first. One problem is in mechanical design, including the mechanical integrity of the membranes and also the membrane-to-metal junctions. Significant consortia efforts formed with US DOE funding led to the development of two novel technologies, known as ITM (ion transport membrane) and OTM (oxygen transport membrane).

Because the production of synthesis gas from natural gas by oxygen-blown ATR accounts for about 60% of the total capital cost of the GTL plant, the ATR and cryogenic ASU are the two most capital-intensive components in the syngas generation process section. ITM/ITM Syngas technology aims to lower capital intensity by combining

the ASU and ATR units into a single reactor. In the ITM/ITM Syngas process, O₂ in air is selectively reduced and transported across the membrane to the other side, where it partially oxidizes methane to form syngas. The ITM/ITM Syngas concept is particularly attractive for use in locations where space for an ASU would be limited, such as on an offshore platform. The ITM Syngas Team, led by Air Products, has estimated capital cost savings of greater than 30% over a conventional ATR with ASU, as well as an increase in overall fuel efficiency to 61% versus 58% for ATR/ASU, with up to 40% less deck space required [22].

OTM oxygen technology aims to lower capital intensity by replacing cryogenic ASU technology with membrane reactors that allow O₂ to selectively pass through the ceramic membrane to the other side, where it partially oxidizes methane to form syngas. The OTM Alliance team, led by Praxair, focused intensively on materials development [23]. This DOE program ended in 2004 and does not appear to be progressing toward commercialization at this time. The OTM approach also has been studied in combination with electrochemical pumping of O₂ [24] and with CO₂ capture using a solid sorbent [25].

Although membrane materials with satisfactory oxygen permeability at the relevant conditions do exist and progress has been made, the level of oxygen flux in all cases still requires improvement. Concerns persist over mechanical integrity and stability with time. Catalytic membrane reactors are still under development, and it is not possible to predict at this time whether or when CMR will be successfully developed for commercial use. The goal for further development of CMR is to avoid the cost of an oxygen plant.

Pre-reforming and Post-reforming

Additional reforming may be performed upstream or downstream of the main reformer unit, when the main furnace unit has limited heat transfer capability that would otherwise limit overall conversion. The pre-reformer reaction breaks down the heavier hydrocarbons (mostly propane and butane) to methane, upstream of the reformer heater. Post-reforming uses an additional catalyst bed downstream of the main reformer. A down-flow reactor is placed between the outlet transfer line and the steam generator that

recovers excess process heat. Adding a post-reformer raises the effective residence time of the reacting gas and thereby boosts capacity. Pre- and, in particular, post-reforming are often considered as retrofits, which may be complicated by having insufficient space or inadequate metallurgy for the higher heat burden.

Methane Reforming for Methanol/DME Synthesis

Synthesis gas production is best characterized by the generalized stoichiometric ratio M (also referred to as the module), where $M = (H_2 - CO_2) / (CO + CO_2)$. Ideally M should be equal to 2.0 for methanol (or for F-T). For kinetic reasons, in order to control by-products formation, a value slightly above 2.0 is normally preferred.

Because about 60% of the capital cost for a methanol plant is in the methane reforming unit, research has aimed at improving reforming processes. Over the years, three process concepts have been considered for preparing syngas targeted for methanol synthesis, described below.

- Single-step tubular reforming
 - Steam/carbon ratio = 2.4.
 - Capacity range up to 2,500 t/day methanol equivalent.
 - Availability of CO_2 increases capacity where single-step tubular reforming is attractive.

In the past, the methanol industry was dominated by single-step, steam methane reforming without the use of oxygen. Today, single-step reforming is considered mainly for relatively small plants and for cases where CO_2 is present in the natural gas or is available on site or nearby from other sources such as an ammonia plant.

- Two-step reforming
 - Steam/carbon ratio = 1.5–1.8.
 - Capacity range about 1,500–7,000 t/day methanol equivalent.
 - Heavy natural gas makes two-step reforming less attractive compared to single-step tubular reforming.
 - Availability of CO_2 is no advantage with two-step reforming.

Two-step reforming is a combination of a fired, tubular reformer and an oxygen-blown secondary reformer. This concept was applied by Statoil and

Conoco in a 2,400 t/day methanol plant in Tjeldbergodden, Norway, commissioned in 1997. The facility has an annual capacity of about 900,000 t of methanol, which corresponds to 25% of Europe's total production capacity [26].

- ATR
 - Steam/carbon ratio = 0.6–0.8.
 - Capacity range about 5,000–10,000 t/day methanol equivalent.
 - Heavy natural gas makes ATR less attractive.
 - Availability of CO_2 is no advantage.
 - This scheme is most attractive for the production of fuel grade methanol.

Using ATR, the composition of synthesis gas is raised to have adequate hydrogen content by taking advantage of the WGSR with removal of CO_2 .

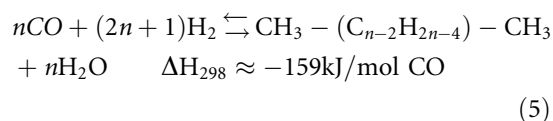
The main conclusions from early studies were that for relatively small plants (<1,500 t/day methanol equivalent), single-step tubular reforming is advantageous for syntheses gas production. For intermediate scale capacities (corresponding to present world-scale capacities, about 1,500–3,500 t/day methanol), two-step reforming is most attractive. For capacities above about 3,500 t/day, ATR or a combination of ATR and SMR is preferred. For example, the Lurgi-combined reforming process uses a combination of steam and autothermal reforming. Proven single-line capacities for the oxygen plants associated with combined reforming have increased to about 3,000 t/day, corresponding to about 7,000 t/day methanol plant equivalent using two-step reforming, or about 5,000 t/day methanol plant equivalent using ATR alone.

A new process concept for the synthesis loop for methanol has been developed and specially adapted to syngas production by ATR at low steam/carbon ratios. In this scheme, CO_2 removal is not applied. Instead, the gas composition is adjusted by recovery and recycle of hydrogen from the purge gas from the synthesis loop.

Fischer-Tropsch Synthesis

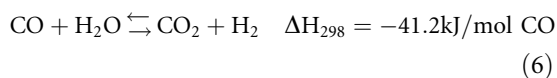
The Fischer-Tropsch (F-T) reaction is a polymerization or oligomerization of carbon monoxide performed catalytically under reducing conditions. As shown in Eq. 5, the F-T reaction is thermodynamically favorable by virtue of high exothermicity. Catalysis, temperature

control, and downstream processing are used to provide a product slate as close to that of target applications as possible.



Fischer–Tropsch processes were originally developed for use in coal-to-liquid (CTL) processes and are similar when applied to the syngas derived from natural gas. The reader is referred to the CTL entry in this publication for additional discussion of F-T performance variables, operating regimes, and product distributions.

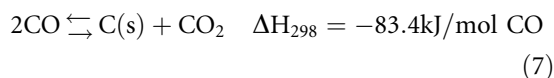
The primary differences between F-T processes applied for GTL and CTL are the range of operating temperatures used, the type of catalyst used, and the type of reactor used. These differences all relate to differing compositions of F-T syngas feed in GTL versus CTL. The stoichiometry of the F-T reaction to make hydrocarbons is about 2.1/1 in H_2/CO as shown in Eq. 5. Syngas produced by gasification of coal has a H_2/CO ratio of 0.7–1.0 and is therefore stoichiometrically deficient in hydrogen. Use of an iron-based catalyst for the F-T step of CTL is typical because Fe concurrently catalyzes the water gas shift (WGS) reaction, providing makeup hydrogen as per Eq. 6.



In contrast, as shown above in Eqs. 2–4, the H_2/CO ratio of syngas produced by reforming methane is at or in excess of the value of 2 required by the stoichiometry of the F-T reaction. Thus iron F-T catalysts are less preferred for use in GTL processes than cobalt F-T catalysts, which do not catalyze the WGS reaction. Cobalt is effective in the low-temperature Fischer–Tropsch (LTFT) regime, operated at about 220°C, and not for high-temperature Fischer–Tropsch (HTFT), operated at about 340°C. At the higher temperatures, Co tends to make methane.

Iron F-T catalysts can be used in GTL processes, performed at low temperature to lessen the contribution of iron-catalyzed WGS reaction. Operating in the LTFT regime, a slurry bed reactor (SBR) is typically preferred for several reasons. One reason is that the

slurry bed uses a much lower catalyst loading than either the fixed bed or fluidized bed reactor technologies used in HTFT processes. A second reason is that the slurry bed can remove reaction heat more rapidly, to avoid temperature increases that would result in undesired formation of high levels of methane, light hydrocarbons, and, in the extreme, catalyst deactivation due to coking, sintering, and disintegration. Deposition of carbon by the Boudouard disproportionation reaction, Eq. 7, is to be avoided, and the heat distribution imparted by the slurry bed is advantageous. A third aspect favoring the slurry bed is that the resultant steam to carbon (S/C) ratio is superior.



Even so, Shell does operate a fixed bed, Co-catalyzed F-T process in its GTL plants in Malaysia and Qatar. Shell commissioned the Bintulu plant prior to the widespread consideration of slurry bubble reactor technology for F-T processes and stands by the fixed bed technology. One advantage of fixed beds over slurry beds is the potential for a lesser degree of catalyst attrition that results from impacts and shear forces in the latter.

Reaction Pathway

Fischer–Tropsch chemistry has long been recognized as in effect a polymerization reaction, wherein the monomeric unit is the methylene group, $-\text{CH}_2-$ that is generated in situ from carbon and hydrogen atoms of CO and H_2 . The basic steps of the F-T reaction are as follows:

1. Adsorption of reactant, CO, on the catalyst surface
2. Chain initiation, by CO association followed by hydrogenation
3. Chain growth, by insertion of additional CO molecules followed by hydrogenation
4. Chain termination, by reductive elimination of paraffin from catalyst surface or by β -hydride elimination of olefin from catalyst surface
5. Desorption of product, from the catalyst surface

When tested under relatively comparable conditions, Fe and Co catalysts give sufficiently similar carbon number distributions to suggest a common intermediate for the different F-T metals. The predominance of linearity

in F-T products and the formation of terminal, α -olefins suggest that the intermediate responsible for chain growth is a single-carbon unit. For at least the past 50 years, most researchers have agreed that the product distribution observed in F-T is consistent with a single intermediate and chain propagation mechanism for all products (with caveats for C_1 and C_2 , vide infra). Research on the mechanism was most defining during the 1960s and 1970s, and experiments and debate continue on the topic of what that single-carbon unit is, and how it derives from the CO or, for iron, the CO or CO_2 present. Candidates include a surface-associated carbon, a metal methylene, a metal hydroxycarbene, a metal alkyl, a metal alkoxide, or a metal carboxylate – a great many choices for a reaction whose mechanism has been under investigation for about a century.

The mechanistic complexity of F-T chemistry is compounded by the use of certain terms, such as carbide or carbene, by different researchers to denote different intermediates. Table 2 shows the steps of the F-T reaction according to various mechanisms currently under consideration in the field, presented as composites that capture the essence of each step. The many descriptors in common use for the mechanistic chemistry and intermediates are also provided.

Formation of C_1 Species The initial reaction of syn-gas with the F-T metal involves reduction of CO and produces a metal-bound C_1 species CH_2 , $CHOH$, or CH_2OH . Metal-carbides have also been considered for the C_1 intermediate, forming by splitting of the C–O bond with formation of water from the O and H_2 ; however, the formation of carbides as discrete species attached with strong, multiple bonds to the surface has been largely discounted by isotopic tracer studies. Many people do still use the term “carbide” to refer to a more loosely associated metal- C_1 species, in what has been referred to as the “modern carbide” mechanism. Surface carbiding of iron has been demonstrated under F-T conditions, but Co does not become carbided unless H_2 is absent. Given that cobalt is a good hydrogenation catalyst, if formed during F-T the carbide would be hydrogenated very quickly.

The intermediacy of immediate C–O bond cleavage and surface carbides need not be invoked if CO itself is hydrogenated prior to C–O bond scission. Devising experiments that would differentiate exactly to what

degree the CO is associated with the metal when scission occurs would be difficult under realistic F-T conditions. In addition, work by Davis has shown that CO_2 can initiate chain growth in Fe catalysis (in which case an initial species would be formic acid) [27].

Chain Initiation and Propagation Chain initiation occurs when the first C–C bond of the chain is formed, by coupling of two initial C_1 fragments or by insertion of carbon monoxide into a metal alkyl bond (or possibly a metal alkoxide bond, in the postulated iron case). Propagation occurs by successive insertion of carbene C_1 species into the growing metal-hydrocarbon chain, or by successive cycles of CO insertion and reduction at the metal. The carbene terminology used here is that used by E.O. Fischer in his pioneering discovery of species of this type and refers to any species $M_1C(A_1)(A_2)$ or $M_2(A_1)(A_2)$, where A_1 is hydrogen or carbon and A_2 is hydrogen, carbon, or oxygen. For monometallic M_1 , the bonding is formalized as $M=C$ and for bimetallic M_2 the bonding is viewed as $M-C-M$. These $M=CH_2$ or $M-CH_2-M$ carbenes are highly reactive and are well known to insert into metal–carbon bonds, as required by the F-T product distribution, under milder conditions and more quickly than carbon monoxide does.

Early studies performed by Pettit with ^{13}C labeling strongly supported the intermediacy of $M=CH_2$ species [28]. Including ^{14}C labeled alcohols and olefins in the feed, Emmett found that enols of alkoxides can initiate chain growth. Much of this early mechanistic work, as well as the concept that CO_2 can initiate chain growth (Fe catalysis), has been reviewed by Davis [29]. The propagation mechanism also can be considered a carbene mechanism, in that the initial product of CO reaction with a metal alkyl is $M-C(=O)R$, which has a tautomeric carbene form, $M^+=C(O^-)R$. Formation of oxygenates could occur by insertion of CO, to give $M-C(=O)alkyl$ intermediates which are then reduced by H_2 . Alternatively the alcohols could form by reductive elimination of metal surface-bound OH and the surface-bound chain.

Chain Termination The typical chain termination reaction is intramolecular, either reductive elimination or β -hydride elimination. Hydrogenation of a metal on which a chain is actively growing allows for reductive

Gas to Liquid Technologies. Table 2 Elements of F-T mechanisms [7]

Formation of C_1 species	Chemistry	Common names for C_1 species
$M_n + CO \rightarrow M_{n-1}-C + M=O \xrightarrow{2H_2} M=CH_2 + (n-1) M + H_2O$	CO dissociation and hydrogenation	Surface-associated carbon (M--C), metal carbide (M-C), methylene (CH_2), carbene ($M=CH_2$), methylidene ($M=CH_2$), alkylidene ($M=CH_2$)
$M + CO \rightarrow M=C=O \xrightarrow{H_2} M=C \begin{array}{c} OH \\ H \end{array}$	Direct CO hydrogenation	Enol, hydroxycarbene
$M-H + CO \rightarrow M-\overset{\overset{O}{\parallel}}{C}-H \xrightarrow{H_2} M-CH_2OH$	CO insertion and hydrogenation	Hydroxymethylene, alcohol
$Fe-H + CO_2 \rightleftharpoons Fe-\overset{\overset{O}{\parallel}}{O}-C-H \xrightarrow{4 Fe-H} Fe-O-CH_3 + H_2O$	CO ₂ insertion and hydrogenation	Oxygenate, metal alkoxide
$2 M=CH_2 \rightarrow \begin{array}{c} H_2C-CH_2 \\ \quad \\ M \quad M \end{array} \xrightarrow{1/2 H_2} \begin{array}{c} H_3C-CH_2 \\ \\ M \end{array} + M \xrightarrow[-H_2O]{CO, 2 H_2} \begin{array}{c} H_3C-CH_2 \\ \\ M \end{array} + M=CH_2$	Carbene coupling	Metal alkyl and methylene (CH_2), carbene ($M=CH_2$), methylidene ($M=CH_2$), alkylidene ($M=CH_2$)
$M=CH_2 \rightarrow \begin{array}{c} H_2C-CH_2 \\ \diagup \quad \diagdown \\ M \end{array} \left(\text{or } \begin{array}{c} H_2C=CH_2 \\ \downarrow \\ M \end{array} \right)$	Carbene coupling (intramolecular)	Metallacycle or olefin complex and methylene (CH_2), carbene ($M=CH_2$), methylidene ($M=CH_2$), alkylidene ($M=CH_2$)
$2 M=C \begin{array}{c} OH \\ H \end{array} \xrightarrow{-H_2O} \begin{array}{c} H_3C-C-OH \\ \parallel \quad \parallel \\ M \quad M \end{array} \xrightarrow{H_2} M=C \begin{array}{c} CH_3 \\ OH \end{array} + M \xrightarrow[-H_2O]{CO, H_2} M=C \begin{array}{c} CH_3 \\ OH \end{array} + M=C \begin{array}{c} OH \\ H \end{array}$	Carbene coupling	Enol or hydroxycarbene
$M-CH_2OH \xrightarrow[-H_2O]{H_2} M-CH_3 \xrightarrow[-H_2O]{CO, H_2} M-CH \begin{array}{c} CH_3 \\ OH \end{array} \xrightarrow[-H_2O]{H_2} M-CH_2CH_3$	CO insertion and hydrogenation	Metal hydroxyalkyl or metal alkyl
$Fe-O-CH_3 \xrightarrow{CO} Fe-\overset{\overset{OCH_3}{\parallel}}{C=O} \rightarrow Fe-\overset{\overset{O}{\parallel}}{O}-C-CH_3$	CO insertion	Oxygenate, metal carboxylate
Chain propagation	Chemistry	Common names for propagation species
$\begin{array}{c} H_3C-CH_2 \\ \\ M \end{array} + M=CH_2 \rightarrow \begin{array}{c} CH_3CH_2-CH_2 \\ \\ M \end{array} + M \xrightarrow[-n H_2O]{n CO, 2n H_2} \begin{array}{c} H_3C-(CH_2)_{n+1} \\ \\ M \end{array} + M=CH_2$	Carbene insertion into metal-alkyl chain	Metal alkyl and methylene (CH_2), carbene ($M=CH_2$), methylidene ($M=CH_2$), alkylidene ($M=CH_2$)

Gas to Liquid Technologies. Table 2 (Continued)

$\begin{array}{c} \text{H}_2\text{C}=\text{CH}_2 \\ \downarrow \text{M} \end{array} \xrightarrow[\text{-H}_2\text{O}]{\text{CO, 2 H}_2} \begin{array}{c} \text{H}_2\text{C}=\text{CH}_2 \\ \downarrow \text{M} \end{array} \xrightarrow{\text{H}_2\text{C}=\text{C}(\text{CH}_3)\text{H}} \begin{array}{c} \text{H}_2\text{C}=\text{C}(\text{CH}_3)\text{H} \\ \downarrow \text{M} \end{array} \xrightarrow[\text{-n H}_2\text{O}]{\text{n CO, 2n H}_2} \begin{array}{c} \text{H}_2\text{C}=\text{C}(\text{CH}_2)_{n-1}\text{CH}_3 \\ \downarrow \text{M} \end{array}$	Carbene insertion into metal-olefin chain	Olefin complex and methylene (CH_2), carbene ($\text{M}=\text{CH}_2$), methyldiene ($\text{M}=\text{CH}_2$), alkylidene ($\text{M}=\text{CH}_2$)
$\begin{array}{c} \text{H}_2\text{C}-\text{CH}_2 \\ \downarrow \text{M} \end{array} \xrightarrow[\text{-H}_2\text{O}]{\text{CO, 2 H}_2} \begin{array}{c} \text{H}_2 \\ \diagup \quad \diagdown \\ \text{H}_2\text{C} \quad \text{CH}_2 \\ \diagdown \quad \diagup \\ \text{M} \end{array} \xrightarrow{\text{H}_2\text{C}-\text{CH}-\text{CH}_3} \begin{array}{c} \text{H}_2\text{C}-\text{CH}-\text{CH}_3 \\ \downarrow \text{M} \end{array} \xrightarrow[\text{-n H}_2\text{O}]{\text{n CO, 2n H}_2} \begin{array}{c} \text{H}_2\text{C}-\text{CH}-(\text{CH}_2)_{n-1}\text{CH}_3 \\ \downarrow \text{M} \end{array}$	Carbene insertion into metallacyclic chain	Metallacycle or olefin complex and methylene (CH_2), carbene ($\text{M}=\text{CH}_2$), methyldiene ($\text{M}=\text{CH}_2$), alkylidene ($\text{M}=\text{CH}_2$)
$\begin{array}{c} \text{H}_3\text{C} \\ \diagup \quad \diagdown \\ \text{M}=\text{C} \quad \text{OH} \end{array} + \begin{array}{c} \text{H} \\ \diagup \quad \diagdown \\ \text{M}=\text{C} \quad \text{OH} \end{array} \xrightarrow{\text{-H}_2\text{O}} \begin{array}{c} \text{H}_3\text{C} \\ \diagup \quad \diagdown \\ \text{M}=\text{C} \quad \text{OH} \end{array} \xrightarrow{\text{H}_2} \begin{array}{c} \text{H}_3\text{C}-\text{CH}_2\text{CH}_3 \\ \downarrow \text{M} \end{array} \xrightarrow[\text{-n H}_2\text{O}]{\text{n CO, 2n H}_2} \begin{array}{c} \text{H}_3\text{C}-(\text{CH}_2)_{n+1}\text{CH}_3 \\ \downarrow \text{M} \end{array}$	Carbene coupling, or carbene insertion into metal-carbene chain	Enol, hydroxycarbene
$\text{M}-\text{CH}_2\text{CH}_3 \xrightarrow{\text{CO}} \begin{array}{c} \text{O} \\ \parallel \\ \text{M}-\text{CH}_2\text{CH}_2\text{CH}_3 \end{array} \xrightarrow[\text{-H}_2\text{O}]{2 \text{ H}_2} \text{M}-\text{CH}_2\text{CH}_2\text{CH}_3 \xrightarrow[\text{-n H}_2\text{O}]{\text{n CO, 2n H}_2} \text{M}-(\text{CH}_2)_{n+2}\text{CH}_3$	CO insertion and hydrogenation	Metal alkyl
$\text{Fe}-\begin{array}{c} \diagup \quad \diagdown \\ \text{O} \quad \text{O} \end{array} \text{C}-\text{CH}_3 \xrightarrow{4 \text{ Fe-H}} \text{Fe}-\text{O}-\text{CH}_2\text{CH}_3 \xrightarrow{\text{CO}} \begin{array}{c} \diagup \quad \diagdown \\ \text{O} \quad \text{O} \end{array} \text{C}-\text{CH}_2\text{CH}_3 \xrightarrow[\text{-n H}_2\text{O}]{\text{n CO, 2n H}_2} \begin{array}{c} \diagup \quad \diagdown \\ \text{O} \quad \text{O} \end{array} \text{C}-(\text{CH}_2)_{n+1}\text{CH}_3$	CO insertion and hydrogenation	Oxygenate, metal carboxylate
Examples of chain termination	Chemistry	Comments
$\begin{array}{c} \text{H}_3\text{C}-(\text{CH}_2)_{n+1} \\ \downarrow \text{M} \end{array} \xrightarrow{1/2 \text{ H}_2} \begin{array}{c} \text{H}_3\text{C}-(\text{CH}_2)_{n+1} \\ \downarrow \text{M-H} \end{array} \longrightarrow \text{M} + \text{CH}_3(\text{CH}_2)_n\text{CH}_3$	Hydrogenation, reductive elimination	
$\begin{array}{c} \text{H}_3\text{C}-(\text{CH}_2)_{n+1} \\ \downarrow \text{M} \end{array} \longrightarrow \begin{array}{c} \text{H}_3\text{C}-(\text{CH}_2)_{n-1} \\ \diagup \quad \diagdown \\ \text{H} \quad \text{C}=\text{CH}_2 \\ \downarrow \quad \downarrow \\ \text{M-H} \quad \text{M} \end{array} \longrightarrow \text{M} + \text{H}_2\text{C}=\text{CH}(\text{CH}_2)_{n-1}\text{CH}_3$	β -elimination, olefin desorption	(The reverse is olefin re-adsorption)
$\text{CH}_3-(\text{CH}_2)_n-\begin{array}{c} \diagup \quad \diagdown \\ \text{C} \quad \text{C} \\ \parallel \quad \parallel \\ \text{M} \quad \text{M} \end{array} \text{OH} \xrightarrow{2 \text{ H}_2} 2 \text{ M} + \left\{ \text{CH}_3(\text{CH}_2)_{n+2}\text{OH} \xrightleftharpoons{-\text{H}_2\text{O}} \text{H}_3\text{C}(\text{CH}_2)_n\text{C}=\text{CH}_2 \right\}$	Hydrogenation, dehydration	
$\begin{array}{c} (\text{CH}_2)_n\text{CH}_3 \\ \diagup \quad \diagdown \\ \text{H}_2\text{C} \quad \text{CH}_2 \\ \diagdown \quad \diagup \\ \text{M} \end{array} \longrightarrow \begin{array}{c} (\text{CH}_2)_n\text{CH}_3 \\ \diagup \quad \diagdown \\ \text{H}_2\text{C}=\text{C} \quad \text{CH}_2 \\ \downarrow \quad \downarrow \\ \text{H-M} \quad \text{M} \end{array} \longrightarrow \begin{array}{c} (\text{CH}_2)_n\text{CH}_3 \\ \diagup \quad \diagdown \\ \text{H}_2\text{C}=\text{C} \quad \text{CH}_3 \\ \downarrow \quad \downarrow \\ \text{M} \end{array} \longrightarrow \text{M} + \text{H}_2\text{C}=\text{C}(\text{CH}_3)(\text{CH}_2)_n\text{CH}_3$	β -elimination, reductive elimination olefin desorption	
$\text{Fe}-\begin{array}{c} \diagup \quad \diagdown \\ \text{O} \quad \text{O} \end{array} \text{C}-(\text{CH}_2)_{n+1}\text{CH}_3 + \text{Fe-H} \longrightarrow \text{Fe} + \begin{array}{c} \text{O} \\ \parallel \\ \text{H}_3\text{C}(\text{CH}_2)_{n+1}-\text{C}-\text{OH} \end{array}$	Hydrogenation	

elimination to give the saturated alkane. Loss of an H atom from the beta-position of the chain (the second carbon away from the metal) allows the formation of a metal-hydride and a π -bound olefin, the latter then dissociating easily from the metal center to give an alpha-olefin. The β -elimination is usually quite reversible, and chain growth may resume if an olefin reinserts into the metal-hydride bond.

ASF Distribution For F-T reactions, the number of moles of species with carbon number n decreases exponentially with n . The weight fraction of product having carbon number, W_n , is given by the Anderson–Schultz–Flory (ASF) distribution (Eq. 8):

$$W_n = (1 - \alpha)^2 n a^{n-1} \quad (8)$$

The parameter a is the Schultz–Flory distribution factor and represents the ratio of the rate of chain propagation to the rate of chain propagation plus the rate of chain termination. The parameter α is the chain growth probability, and $1 - \alpha$ is the probability of chain termination. Both parameter a and parameter α are assumed to be independent of chain length. Either value can be obtained graphically using Eq. 8 in log form, with analytical data for W_n as obtained, for example, using gas chromatography. The α values for oxygenated products appear to be about the same as those for hydrocarbons, which suggests that the products all form by the same or similar chain propagation steps and with similar rates of chain termination, albeit with different chain termination chemistries for the oxygenates, olefins, and paraffins.

The parameter α is used most often to characterize product distribution. Values of $\alpha > 0.9$ are representative of processes producing high amounts of wax, α values of about 0.8 maximize diesel cuts, and α values of about 0.7 are characteristic of naphtha production.

F-T Catalysts

Early studies reported that the level of reaction activities of metal catalysts for the Fischer–Tropsch reaction decline in the order Ru, Ni, Co, and Fe. Later studies indicated that activities declined in the order Ru, Fe, Ni, and Co with alumina as support, and in the order Co, Fe, Ru, and Ni with silica as support. In practice, Ru gives rapid production of high molecular weight wax

but this metal is cost-preclusive and supply-limited at plant scale. Nickel gives predominantly methane as product.

Commercially, Co appears to be more active than Fe in that the turnover rates on Co catalysts are about three times higher than on Fe catalysts. In general the observed rates depend on the diffusion rates of reactants and products migrating in and out of the porous catalyst particles, as well as on the intrinsic rate of the F-T reaction at the catalyst surface. The diffusion rates in turn depend on the catalyst's porosity, pore sizes, and particle size, and on whether or not liquid wax is present within the catalyst particles.

The product distribution can be significantly influenced the through use of promoters. The ability of numerous promoters to increase activity with Co catalysts is well documented but their ability to affect selectivity is less clear. In general, any promoter metal that can enhance the even distribution of reduced cobalt on the internal and external surfaces of the catalyst particles could increase the number of active sites available for chain growth and thereby increase the selectivity for wax. Ruthenium appears to do this best, and lanthanum also helps. Rhenium is sometimes used to promote chain growth with cobalt although reports of its efficacy vary.

The traditional industrial support materials for cobalt are alumina, silica, and titania [30]. The choice of support material appears to have little effect on selectivity; however, catalysts using mesoporous supports such as MCM-41 have provided somewhat increased C_{5+} selectivity [31].

The performance of iron catalysts is very dependent on the addition of strong alkali promoters. Oxides of the alkali metals, particularly potassium, promote formation of wax in LTFT catalysis by iron. Use of K_2O also helps keep the iron reduced in carbided form, which may promote the growth of longer chains. Not only is selectivity markedly dependent on the amount of alkali present, but alkali promoters also have a direct effect on the overall rate of reaction. For precipitated iron catalysts, used for LTFT, catalyst activity decreases when the alkali content is raised beyond a threshold level.

Cobalt catalysts produce considerably less CO_2 during F-T synthesis than do iron catalysts.

The partial pressures of water and CO₂ affect selectivity, with the pressure of CO₂ being more relevant for iron than cobalt. Cobalt is a better hydrogenation catalyst than iron, and the products of its F-T reaction tend to have a lower olefin/paraffin ratio than those of iron.

Catalyst Beds and Reactors for F-T Synthesis

Axial temperature gradients, poor heat distribution with carbon deposition, and scale-up limitation mean that multi-tubular reactors are usually not suitable for commercial LTFT synthesis. The slurry bed reactor (SBR) is an attractive alternative to fixed bed reactors. When using precipitated catalysts of identical composition, a SBR system is as active as a fixed bed system despite the much lower catalyst loading of the SBR. Slurry bed reactors have become the technology of choice for companies planning new GTL plants, with the exception of Shell as noted above. Both Sasol and Exxon have operated large demonstration slurry phase F-T units with supported cobalt catalysts. A drawback of SBR technology to date is that additional equipment is required to achieve near complete separation of the finely divided catalyst from liquid wax in slurry systems.

Multi-tubular fixed beds do have advantages as research reactors, because they are easy to operate. No equipment is required to separate the heavy wax product from the catalyst, since the liquid wax simply trickles down the bed and is collected in a downstream knock-out pot. Alternatively, laboratory studies may be performed in high velocity, high recycle stirred tank reactors. By varying the fresh feed gas flow to these units, the partial pressures and the kinetics at different conversion levels (tantamount to different bed depths) can be determined.

Kinetics

Various kinetic equations based on research with laboratory reactors have been generated for the F-T reaction. Equations 9–14 are representative examples of kinetics proposed for Co systems.

$$r = kP_{H_2}^2/P_{CO} \quad (9)$$

$$r = aP_{CO}P_{H_2}/(1+bP_{CO})^2 \quad (10)$$

$$r = aP_{CO}P_{H_2O}^{0.5}/(1 + bP_{CO} + cP_{H_2O}^{0.5})^2 \quad (11)$$

$$r = aP_{CO}P_{H_2}^2/(1 + bP_{CO}P_{H_2}^2) \quad (12)$$

$$r = c(P_{H_2O}^{0.5}/P_{CO_2})/(1 + 0.93P_{H_2}/P_{H_2O}) \quad (13)$$

$$r = d(P_{H_2}P_{CO})/P_{CO} + 0.27P_{H_2O} \quad (14)$$

Equations 15–18 are representative of kinetics proposed for Fe systems.

$$r = kP_{H_2}^{0.6}P_{CO}^{0.4} - fr^{0.5}P_{H_2O}^{0.5} \quad (15)$$

$$r = mP_{H_2}P_{CO}/(P_{CO} + nP_{H_2O}) \quad (16)$$

$$r = aP_{CO}P_{H_2}^2/(P_{CO}P_{H_2} + bP_{H_2O}) \quad (17)$$

$$r = aP_{CO}P_{H_2}/(P_{CO} + bP_{H_2O} + cP_{CO_2}) \quad (18)$$

Equation 10 for Co responds to an increase in pressure: as the pressure increases, the F-T reaction rate decreases. Cobalt is much more resistant to oxidation than iron, whether by oxygen or by water, which could account for the absence of a P_{H_2O} term in kinetic equations for cobalt catalysts. Even so, Co crystals smaller than a certain size (<5 μm) may be oxidized by water vapor and therefore a more detailed kinetic equation for cobalt could include the partial pressure of water.

The kinetics of Fe catalysis are better studied than those for Co. Sasol performed a large number of experiments and measurements in both fixed and fluidized bed reactors to study iron-based catalyst activity in commercial HTFT and LTFT processes. The key findings are as follows:

- Reaction rate is strongly dependent on hydrogen partial pressure. At low conversion levels, reaction rate is solely dependent on hydrogen partial pressure.
- Reaction rate increases with the partial pressure of CO.
- The partial pressure of CO₂ does not appear to have a direct effect on reaction rate. It can, however, affect the partial pressures of other components via the water gas shift reaction and thus may have an indirect effect.
- Reaction rate is markedly depressed by the partial pressure of water.
- The level of hydrocarbon products in the reactor has no apparent effect on F-T reaction rate.

Based on these observations, a satisfactory rate equation for iron-based catalysts is given by Eq. 19:

$$r = mP_{H_2}P_{CO}/(P_{CO} + aP_{H_2O}), \quad (19)$$

where $a = k_{H_2O}/k_{CO}$

The above kinetic expression is speculative but has been shown to estimate the overall conversion for both HTFT and LTFT reactors. This kinetic equation also predicts that if the total pressure of the gas feed is increased, the rate is increased by the same factor, that is, the residence time in the catalyst bed remains the same and the degree of conversion remains unchanged. This means that the production rate is increased in proportion to the increase in pressure. Pilot plant tests at Sasol for both the fixed bed LTFT and the fluidized bed HTFT operations confirmed these kinetic predictions.

Given that reforming methane gives as much or more H_2 as needed for the F-T reaction, the presence of water gas shift chemistry in the F-T reactor could easily become a liability to efficiency of a GTL process. Using the same reactor under the same operating conditions and feed gas flow, a Co catalyst system will have a much higher syngas conversion than an iron catalyst. With an iron catalyst, after water has been condensed out a larger portion of the tail gas is recycled to the reactor. This means that more reactors are required and there is also a greater need to recycle, both of which increase the total fixed capital and operating costs.

Commercial Activities in the Fischer–Tropsch Synthesis

In 1993, two large-scale F-T plants were built specifically for operation with natural gas (as opposed to coal). Backed with funding from the South African government, Mossgas (now PetroSA) began operating a 22,000 bbl/day plant using Sasol technology. The gas is produced from a South African offshore field. Also that year, Shell built a 10,000 bbl/day plant at Bintulu, Malaysia. The \$600 million plant was expanded to 12,500 bbl/day in May 2000, after a fire destroyed the oxygen plant. Also in 1993, a plant designed by Rentech was built in Colorado Springs, Colorado for conversion of landfill gas to liquids. A lack of methane from the landfill shut down the plant.

A joint venture of Qatar Petroleum and Qatar Shell has recently built a GTL plant in Ras Laffan Industrial City, Qatar, using Shells' cobalt-catalyzed, fixed bed technology. The project is referred to as Pearl and is scheduled for the latter part of 2011. At \$18+ billion, the cost of Pearl significantly exceeded budget. This may have resulted at least in part from the incorporation of an additional facility, for producing NGL, into the design. Pearl is expected to generate a total of 140,000 bbl/day in two trains operating at full nameplate capacity, making this plant the world's largest GTL project [32]. In addition Pearl has a capacity for 120,000 barrel of oil equivalents total per day of NGL from two trains. Another new GTL plant has been under construction recently in Qatar, by Sasol, and was operating at about 55% of nameplate capacity in 2010.

F-T Product Upgrading

Conversion of syngas through the F-T process leads to a distribution of products essentially consisting of *n*-paraffins (>90%) and smaller percentages of alcohols, olefins, and methane. Typical hydrocarbon product cut weight fractions are shown in Table 3.

A consequence of aliphatic chain growth by one-carbon units is the theoretical impossibility of synthesizing a F-T product having a narrow range of chain length. This wide molecular weight range of the raw F-T product slate means that downstream separations and refining must be used to create fractions for specific end uses. The crude product from the F-T reaction is most commonly separated into a light, relatively low boiling condensate fraction and a heavy, high boiling

Gas to Liquid Technologies. Table 3 Hydrocarbon weight fractions in typical F-T product

Product cut	Carbon No.	wt%
Gas	C ₁ –C ₄	2.5
Naphtha	C ₅ –C ₉	6.7
Diesel	C ₁₀ –C ₁₉	18.5
Soft wax	C ₂₀ –C ₃₄	27.3
Hard wax	C ₃₅ +	45.0

wax fraction. A typical boiling range cut temperature would be 372°C. The condensate fraction is a liquid that can be shipped from a remote area to a refinery site for upgrading. Condensate containing, for example, C₄–C₁₅ products is further separated, into cold separator liquid and hot separator liquid. The cold separator liquid contains C₅₊ in the boiling range of <260°C, retains about 95% of the total oxygenates produced, and has sufficiently low olefin concentration that hydrogenation is unnecessary. The hot separator liquid contains mostly hydrocarbons and has a boiling range of 260–372°C.

Hydrocracking Heavy Paraffins (Wax)

The C₁₆₊ reactor wax may be upgraded by various hydroconversion reactions including hydrocracking, hydroisomerization, catalytic dewaxing, isodewaxing, or various combinations thereof, to produce middle distillates for use as fuels, lubricants, chemicals, and specialty materials such as nontoxic drilling oils, technical and medicinal grade white oils, chemical raw materials monomers, polymers, emulsions, isoparaffinic solvents, and other specialty products. For example, wax from F-T synthesis is typically first hydrofined, to eliminate alkenes and oxygenated compounds, and then fractionated into different grades of product waxes. The waxes can be sold directly or hydroprocessed by recycle to extinction to yield high quality diesel and kerosene fuels.

Hydrocracking is the most relevant upgrading process, including for maximum production of diesel fuels. Hydrocracking heavy paraffins serves to lower the boiling range of product from that of wax to that of middle distillates, and to improve the cold flow properties of F-T diesel with branched molecules formed during hydrocracking. Because F-T products are predominantly linear, especially those from LTFT, the raw middle distillates have very poor cold flow properties that would hamper their use as transportation fuel unless subjected to hydrocracking.

Another property affected by hydrocracking is the cetane number of diesel. As the pressure of a LTFT reactor increases, the selectivity for wax increases, the degree of branching decreases, and consequently the cetane number of downstream diesel cuts decreases. Subsequent selective hydrocracking of crude wax makes the largest contribution to the final diesel

fuel pool. Some chain branching occurs during hydrocracking and thus the cetane number of the diesel fuel produced is somewhat lower than that of straight run F-T diesel. The final diesel pool, however, has a cetane number above 70 – significantly superior to the cetane number for refinery naphtha, which varies from 40 to 50.

Upgrading wax by hydrocracking to middle distillates gives naphtha as the main coproduct. The naphtha fraction may be upgraded to gasoline, via catalytic reforming, or used as a steam cracker feedstock for olefins production. Converting these two naphtha cuts into on-specification gasoline would require a considerable amount of further octane number upgrading. However, this naphtha consists almost entirely of linear alkanes and therefore makes excellent feedstock for production of ethylene by steam cracking, giving much higher selectivity for ethylene than that obtained from steam cracking of naphtha obtained from crude oil.

Feeds derived from LTFT can be hydrocracked under much milder conditions than typical feeds such as vacuum gas oils that are derived from crude oil. In the hydrocracking of crude oil-derived feeds, typically pressures as high as 150 atm are required to prevent coking of the catalyst by aromatic compounds. This is not necessary with paraffinic feeds, wherein LTFT products are hydrocracked at pressures from 35 to 70 atm using commercial hydrocracking catalysts. The relatively low levels of oxygenates, mainly alcohols and lesser amounts of acids or carbonyls, present in F-T waxes are easily and completely hydrodeoxygenated.

The refining of syncrude is often performed in a separate facility from the upstream GTL process. For direct on-site F-T reaction product upgrading, however, the hot separator liquid cut may be combined with F-T wax and processed under much milder conditions than typical crude oil-derived feeds. Conventional hydrocracking of refinery feedstocks is performed in a fixed bed reactor using multiple or single Co/Mo catalyst compositions under operating pressures as high as 2,500 psi, using conventional technologies such as those offered for license by Chevron or UOP. The high operating pressure is required to prevent coking of the catalyst by aromatic compounds contained in the feed. Such high pressures are not

required when hydroprocessing paraffinic feeds from LTFT synthesis, which do not contain aromatics, and lower pressures between 500 and 1,000 psi are used to hydrocrack LTFT products with commercial hydrocracking catalysts. During this process, the oxygenates present in the feed are easily and completely hydrodeoxygenated. Typical hydrocracking conditions for F-T products entail operating temperatures of 320–450°C, pressures of 1,000–1,500 psi, and hydrogen treat rates of 500–5,000 scf/bbl.

The low sulfur content of syngas from natural gas enables consideration of hydrocracking catalysts containing a noble metal component, particularly platinum, instead of the less expensive, conventional sulfur-resistant Co/Mo catalysts. Platinum leads to higher hydroisomerization activity and less severe hydroprocessing conditions, with better low-temperature performance of the resultant diesel fuel. For example, ExxonMobil's hydroisomerization process uses a noble metal catalyst to increase the amount of isoparaffins in the distillate fuel and to help the fuel meet pour point and cloud point specifications without the need for additives. Hydroisomerization may be carried out with a single Pt-based catalyst in a fixed bed reactor. Preferred operating conditions use a temperature range of 260–400°C, a pressure range between 500 and 1,000 psi, and a hydrogen treat rate of 1,000–4,000 scf/bbl, and heavy wax products are recycled to extinction through the hydrocracking process.

The processing pressure is dependent on the hydrogenation capacity of the catalyst. When the hydrogen partial pressure is too low, dehydrocyclization of the paraffins starts to occur with the formation of polynuclear aromatics, which eventually would lead to deactivation of the catalyst due to coking. Lower hydrogen/wax ratios lead to a decrease of conversion rate of the C₂₂₊ fraction and, after adjustment of reaction conditions to achieve the same conversion levels, an increase of both the isoparaffin content and the selectivity to products lighter than diesel.

Isoparaffin content also increases with operating pressure. The hydrocracking process has to meet the following conditions:

- The chain length of the hydrocracked fragments should be predominantly in the desired product's carbon number range.

- The components above said desired range should be hydrocracked in preference to those that are already within or below the desired range.
- The production of less commercially attractive species, for example, light hydrocarbon gases, should be minimized.

Due to the clean nature of the feed, non-sulfided hydrocracking catalysts containing a noble metal component such as platinum can be considered. The use of noble metal catalysts leads to higher hydroisomerization activity and consequently better low-temperature characteristics for the diesel product. Most probably all C₁₄₊ material would be treated using such catalysts, due to the enhancement of cold flow properties resulting from simultaneous isomerization and hydrocracking. A high blending cetane value is retained, and very little of the diesel range material in the feed is degraded into naphtha and kerosene. Catalysis using a noble metal on amorphous silica-alumina is likely to be the preferred approach for maximum production of diesel blending material.

There are many technology licensors for hydroconversion processes, including ChevronTexaco, UOP, IFP (Axens), Criterion, and Haldor Topsøe. Catalysts can be obtained from these companies as well as from Akzo Nobel and Sud-Chemie, among others.

Hydrotreating Light Paraffins

Hydrogenation is less costly than hydrocracking. For large-scale plants, a condensate hydrotreater should be used to process hydrocarbons that are already lighter than the desired middle distillates and also for straight run middle distillates if cold flow properties are not an issue. A condensate fractionator may be employed if it is desired to produce linear paraffin products. The condensate fractionator can also serve to remove light gases in order to stabilize the condensate for intermediate storage at atmospheric pressure. If a nearby naphtha cracker is available, then the condensate fractionator will provide a naphtha product that need not be hydrogenated to ensure storage stability. Also, a kerosene fraction can be separated using the condensate fractionator.

If both hydrotreating and hydrocracking are used, they may be operated at the same pressure and use a common hydrogen loop. Alternatively, the

hydrotreater can be operated in a once through mode, at lower pressure, and the hydrogen tail gas then subsequently pressurized and fed to the hydrocracker loop. For a 30,000 bbl/day equivalent product upgrading system comprised of hydrocracking and hydrotreating, the size of the unit is comparable to those used in oil refineries. There are many process configurations possible depending on site-specific factors like plant capacity, available land, and possible integration with other facilities.

While product upgrading schemes center around three fuel products (diesel, LPG, and naphtha), the upgrading plant is quite amenable to different configurations if the market demand so indicates. Even with cracking and hydrotreating, the ASF distribution means that the F-T process and its associated refinery produce other transportation fuel and products beyond middle distillates. Other saleable products may include solvents, kerosene, jet fuel, illuminating paraffins (sold for use in lighting and appliances in South Africa), and base oils (for lubricant products). Maximizing the recovery of the higher boiling jet fuel while maintaining diesel production will tend to increase economic return.

LTFT Naphtha and Blends

Four LTFT naphthas can be produced using process configurations summarized below in Table 4.

The slurry phase distillate (SPD) naphtha is an excellent feed for the production of lower olefins, particularly ethylene, by steam cracking. Such paraffinic naphtha feed demands less feedstock and lower energy consumption in the cracking process than naphthas from other sources. The optimum olefin yields and

selectivities obtainable from any naphtha depend on the cracking severity at commercial conditions. The mass ratio of light olefins, propylene/ethylene (P/E), has an inverse relationship to the cracking severity. High P/E values correspond to low severities and also lower coproduction of methane.

The LTFT naphthas other than SR are essentially entirely paraffinic and therefore have compositions far from those of high-octane naphthas useful in gasoline blends. Even so, the conventional naphthas typically require that the refiner improve cold flow characteristics by changing the boiling point, for compatibility with low-temperature operation. Blends of SPD naphtha with conventional diesels could be used to meet winter grade fuel specifications without decreasing the production volume of diesel and other light products obtained from oil refining. Including F-T naphtha in blends can assist refiners in meeting low sulfur specifications in fuels that have sulfur content near zero (below 1 ppm).

The projected threshold for economic viability of F-T GTL technology, in terms of crude oil price, is now at about \$70–75 per BOE. This threshold is based on recycle of reformed NGL (vs. production of NGL for market sales). The investment cost for a modest scale GTL facility producing predominately diesel fuel has been estimated at about \$1.1 billion [7]. There are then further opportunities to add value by producing base oils and waxes.

This \$1 billion investment value may be taken as a starting point when estimating actual investment costs for a first-of-a-kind plant. There is a substantial gap between such estimate and the final cost of, for example, Pearl. In the Pearl project, the initial configuration was changed to one utilizing NGL for market sales and also the builders experienced escalating capital costs, due to a shortage of major equipment, procurement delays, shortage of skilled engineering labor, and an increased number of reactors from an initially conservative estimate, among other contributing cost factors.

Methanol

General Process Description

Currently the majority of methanol is synthesized from syngas produced from natural gas, using autothermal reforming, steam reforming, or a combination of steam reforming and autothermal reforming. Most existing

Gas to Liquid Technologies. Table 4 Process schemes for production of LTFT naphtha

F-T naphtha		Production scheme
SR	Straight run	Fractionation of F-T condensate
HT SR	Hydrogenated straight run	Fractionating and hydrotreating of F-T condensate
HX	Hydrocracked	Fractionating and hydrocracking of F-T wax
SPD	Slurry phase distillate	Blend of HT SR and HX naphthas

commercial plants use steam reforming. Incoming pressurized natural gas feed is preheated and passed through a zinc oxide catalyst bed to remove sulfur compounds that would otherwise poison downstream catalysts. Steam is added to the natural gas stream before it enters the reformer. A typical reformer is a large high-temperature furnace containing a number of vertical tubes that are filled with a nickel catalyst. Once the natural gas is reformed, the resulting synthesis gas is cooled to about 30°C, condensed water is separated from the gas in a process gas separator, and the syngas is compressed using a steam turbine compressor that uses steam generated in an economizer on the syngas outlet from the reformer.

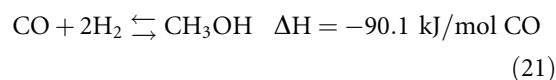
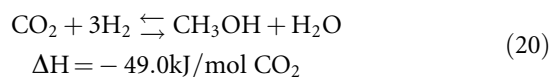
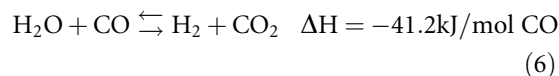
Compressed syngas containing H₂, CO, and CO₂ is fed to a methanol synthesis reactor that uses a copper catalyst. The methanol reactor outlet gas is cooled to about 40°C and then passed to a pressure letdown tank where unreacted synthesis gas is separated from the product methanol and water. A portion of this unreacted syngas stream is purged and burned as fuel, while the majority is compressed by a recirculating compressor and remixed with feed synthesis gas (from the upstream syngas compressor). The unreacted syngas is thus recirculated back to the methanol reactor, resulting in an overall conversion efficiency of 99%.

Crude methanol collected in the letdown tank usually contains up to 18% water plus impurities such as methyl formate, higher alcohols, ketones, ethers, and esters. This is stored at atmospheric pressure in an intermediate storage tank and later fed to a distillation plant. The distillation plant may be configured in a variety of ways. A typical distillation plant consists of a single distillation column that removes the light ends including dissolved gases, dimethyl ether, methyl formate, and acetone that are burned as fuel. Bottoms from the light ends column are sent to a system that consists of two distillation columns, both of which deliver pure methanol overhead and sequentially remove water and heavy ends such as higher alcohols, ketones, and esters (formed by esterification of lower alcohols with formic, acetic, and propionic acids in column bottoms). Water from the bottoms of the first distillation column is discharged to the process sewer system, while heavier ends separated as bottoms in the second methanol recovery column are sent to the reformer furnace fuel system for

combustion. The methanol is at least 99.8% pure after this distillation and is transferred from temporary storage to final product storage tanks where it is directly pumped into ships, railcars, or trucks for delivery to customers and distributors.

Reaction Pathway

Chemical reactions considered in methanol synthesis are shown in Eqs. 6, 20, and 21.



Both CO and CO₂ are reactive in the synthesis. Coupling the methanol synthesis with the water gas shift reaction (Eq. 6) allows a larger quantity of carbon to become available for most efficient synthesis. Although the direct hydrogenation of either CO₂ or CO to methanol is thermodynamically favorable (Eqs. 20 and 21, respectively), the activation energy for direct hydrogenation is apparently too high and instead the reaction proceeds through a different mechanistic pathway.

The reaction of syngas to produce methanol is about 100 times slower if CO₂ is not also present in the mix. Until as recently as the 1990s, the role of CO₂ in methanol synthesis was unclear. The water gas shift activity of Cu catalysts is so high that it was difficult to delineate the roles of CO and CO₂ in methanol synthesis. Isotopic labeling studies have now unequivocally shown that CO₂ is the source of carbon in methanol. The CO undergoes the WGS reaction to make H₂ and CO₂. The CO₂ is thought to keep the catalyst in an intermediate oxidation state, thus preventing the reduction of ZnO followed by the formation of brass.

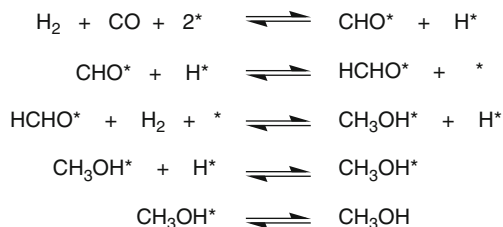
A long-held mechanism for catalytic methanol synthesis proceeds through a metal-bound formate intermediate, formed by hydrogenation [33, 34]. As depicted in Eqs. 20–24, CO₂ is adsorbed onto the catalyst surface [35], where it reacts with a partially oxidized methane intermediate to form a carbonate, which is then hydrogenated. This intermediate is then

further hydrogenated in the rate-limiting step. The copper catalyst sites have high activity for splitting the first C–O bond of CO₂ that helps maintain the oxidation state of the active copper sites. At high concentrations, however, CO₂ actually decreases catalyst activity and inhibits methanol synthesis. The feed gas composition for methanol synthesis is typically adjusted to contain 4–8% CO₂ for maximum activity and selectivity. Although Cu has water gas shift activity, excessive water causes deactivation by blocking active sites. Site blocking lowers catalyst activity but improves selectivity, with a 50% decrease in by-product formation.

Recent modeling efforts using density functional theory (DFT) suggest that the traditional methanol synthesis mechanism proceeding via CO₂ accounts for only about two third of the methanol produced [36]. The remaining methanol is said to derive from the CO, mostly via direct reduction by H₂. The proposed mechanism is conceptually similar to that for CO₂ reduction to methanol and is shown in Scheme 1. Overall, methanol synthesis appears rate limited by formation of methoxy (CH₃O*, where * refers to the catalyst surface) at low CO₂/(CO + CO₂) ratios and by CH₃O* hydrogenation in CO₂-rich feeds. Hydrogenation of CH₃O* as per Eqs. 20 and 21 is the slow step for both the CO and the CO₂ methanol synthesis routes.

Reactors and Catalysis for Methanol Synthesis

The two main process features considered when designing a methanol synthesis reactor are (a) controlling and dissipating the large reaction heat, and (b) overcoming the equilibrium constraint to maximize per pass conversion efficiency. Methanol synthesis is a version of the classic trade-off between conversion and selectivity. The maximum per pass conversion



Gas to Liquid Technologies. Scheme 1

Proposed mechanism for reduction of CO to methanol

efficiency of syngas to methanol is limited to about 25%, and catalyst activities generally decrease as the temperature is lowered. At higher temperatures, catalyst activity increases but so does the chance for competing side reactions to form by-products such as dimethyl ether, methyl formate, higher alcohols, and acetone. Catalyst lifetimes are also reduced by continuous high-temperature operation.

Maintaining reaction temperatures below 300°C is the trade-off strategy employed to minimize catalyst sintering while maximizing conversion of syngas to the target product slate. Although the equilibrium conditions favor low temperatures, methanol converters must be operated at temperatures in the range 200–300°C to ensure the catalysts are active and to use the heat of reaction effectively [37].

Removing methanol as it forms is another strategy used for overcoming the thermodynamic limitations and improving the per pass conversion and process efficiency. Methanol is either physically removed or converted to another product, such as dimethyl ether, methyl formate, or acetic acid.

Numerous methanol reactor designs have been commercialized over the years and these can be roughly separated into two categories, namely, adiabatic and isothermal reactors. Adiabatic reactors often contain multiple catalyst beds separated by gas cooling devices, using either direct heat exchange or injection of cooled syngas, either fresh or recycled. Axial temperature profiles often have a saw tooth pattern: dropping down at the point of heat removal and increasing linearly between the heat exchange sections. The isothermal reactors are designed to continuously remove the heat of reaction so they can operate essentially like a heat exchanger with an isothermal axial temperature profile.

Currently all commercial processes produce methanol from syngas in a vapor phase reactor with a synthesis loop. In a gas phase, tubular or packed bed reactor, reaction heat is dissipated in situ by injecting cool, unreacted feed at various points along the reactor length, or by including internal cooling surfaces for heat exchange. Temperature moderation is also achieved by recycling large quantities of gas, with gas cooling in the recycle loop. Even with these heat transfer measures, the CO content of the feed is limited to about 16% as a means of controlling heat content by limiting the conversion.

The design challenge of removing heat while maintaining precise temperature control is significant, because excessive temperatures seriously diminish the lifetime of the catalyst. A liquid process is of interest, wherein nonvolatile mineral oils make up part of the reactor liquid phase and act as a temperature moderator and heat sink. A new approach investigated by Air Products involves bubbling the syngas through a slurry consisting of micrometer-sized methanol synthesis catalyst and mineral oil as reaction medium that transfers heat from the catalyst surface, through the slurry, to boiling water in an internal tubular heat exchanger. Concentrations of CO as high as 50% have been used in the slurry system without damaging the catalyst activity.

A typical commercial methanol synthesis catalyst contains of copper and zinc oxide phases with alumina support, that is, Cu/ZnO/Al₂O₃. Recent modifications to catalyst preparation at laboratory scale include specialized coprecipitation methods [38], for a greater synergistic effect between copper and zinc, and the maintaining of a strict control of pH and the temperatures of precipitation and calcination [34]. As shown in Table 5 and Fig. 3, the productivity of commercial catalysts is sensitive to several factors, including conditions used for preconditioning by, for example, aging the catalyst [39, 40].

Activities in Methanol Synthesis

Catalytic methanol synthesis from syngas was classically run as a high temperature, high pressure, exothermic, equilibrium-limited synthesis reaction operated at 320–450°C and 250–350 atm pressure. BASF of Germany introduced the first commercial synthesis of methanol, in 1923, using a zinc/chromium oxide

catalyst. New methanol synthesis technologies based on lower reaction pressures (50–100 atm) were introduced in the 1960s. Whereas high-pressure processes require 139,000–146,000 Btu/gal of methanol produced, all modern plants are based on low-pressure technology and are expected to require fewer than 100,000 Btu/gal of methanol produced, on an average yearly basis. Additional advantages of low-pressure processes include lower maintenance costs and reduced downtime.

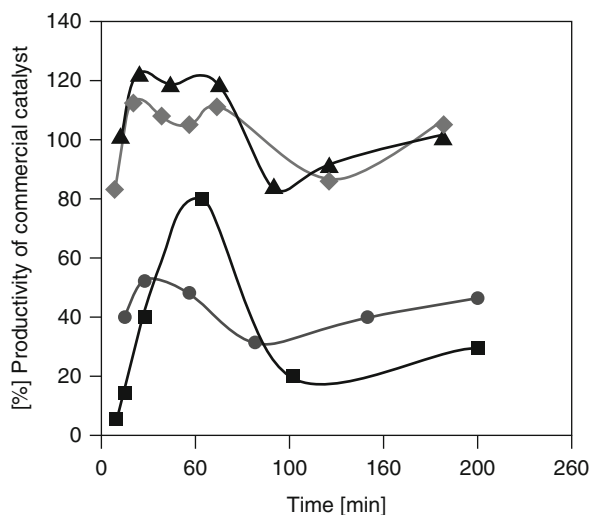
Johnson Matthey (formerly ICI Chemical Industries) of the United Kingdom and Lurgi of Germany offer the two most widely used technologies for methanol synthesis. The ICI low-pressure process operates at pressures between 50 and 100 atm and temperatures of about 210–290°C. A highly selective copper/zinc alumina catalyst is used in a shot-cooled single or multi-bed reactor. The reaction is quenched on the bed by introducing cold gas through specially designed spargers. Heat recovered from the synthesis loop provides part of the process steam required by the steam reforming furnace.

Lurgi's conventional low-pressure methanol process has a maximum output of 1,830–2,000 t/day. It uses a copper-based catalyst, pressures of 50–100 atm, and temperatures of 230–265°C. Temperature control of the catalyst is achieved by indirect cooling with boiling water. The boiling water in the reactor shell generates 1–1.2 kg of steam/kg of methanol, with a pressure of up to 45 atm. The steam, after moderate superheating, supplies the total energy for driving the recycle gas compressor and is subsequently used for the reboilers in the methanol distillation section. A small quantity of surplus steam is often available for export.

Johnson Matthey, Lurgi, and Haldor Topsøe have been pursuing 10,000 t/day mega-methanol single

Gas to Liquid Technologies. Table 5 Methanol synthesis catalysts prepared under various aging conditions

Catalyst	Metals (molar ratio)	Precipitation parameters	Aging parameters
Cat _{2-const}	Cu/Zn/Al (60/30/10)	pH 7, 70°C	pH 7 _{const} , 70°C
Cat ₃	Cu/Zn/Al (60/30/10)	pH 8, 70°C	pH 8 _{const} , 70°C
Cat _{2-free}	Cu/Zn/Al (60/30/10)	pH 7, 70°C	Free-aged, 70°
Cat _{Cu/Zn}	Cu/Zn (70/30)	pH 7, 70°C	pH 7 _{const} , 70°C



Gas to Liquid Technologies. Figure 3

Methanol productivity as a function of the aging time for selected catalysts: (triangle) $\text{Cat}_{2\text{-consti}}$; (square) $\text{Cat}_{2\text{-freei}}$; (circle) $\text{Cat}_{\text{Cu/Zn}}$; (diamond) Cat_3 [39]

train plant designs. Lurgi's MegaMethanol[®] process can accommodate a single train capacity of at least 5,000 t/day and was developed for capacities larger than one million tons of methanol per annum. A schematic of Lurgi's process is shown in Fig. 4. Lurgi does not use conventional steam reforming to produce syngas, instead employing autothermal reforming with a newer burner system in what is called the MegaSyn process for converting oxygen and natural gas to syngas.

The first plant to use the MegaMethanol[®] technology was brought on stream in 2004 at Point Lisas in Trinidad. The plant is owned by Atlas Methanol Company, a joint venture between Canada's Methanex and BP Trinidad and Tobago (bpTT). The 1.7 t/year Atlas methanol plant was built adjacent to the existing Titan Plant, which has a methanol production capacity of 850,000 t/year. Average total methanol production from the plants is 2.5 million tons per year, which makes Trinidad and Tobago the world's leading methanol exporter.

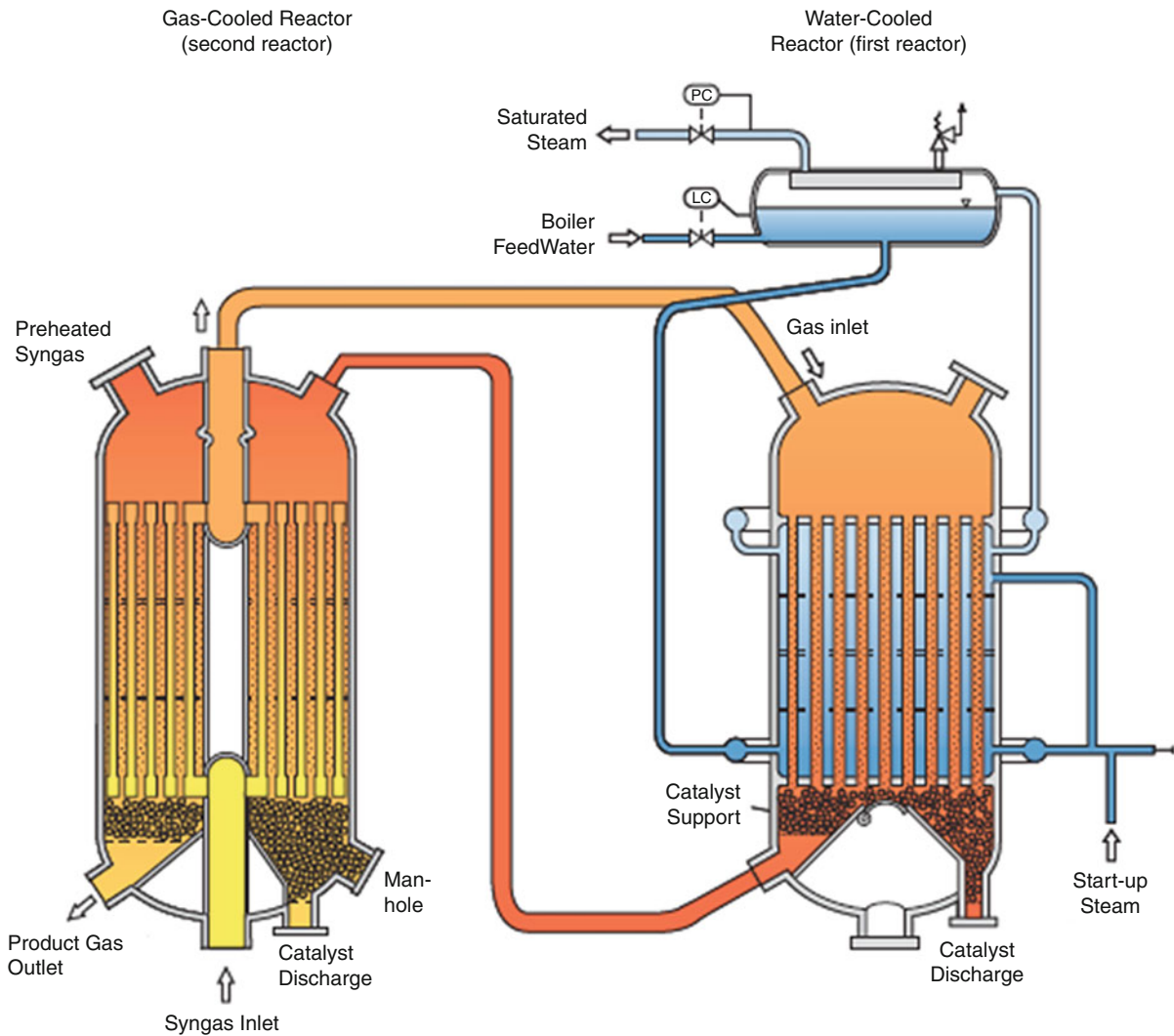
The Point Lisas plant has an on-site air separation plant having a production capacity of 2,800 t of pure oxygen per day. Their mega-methanol production process consists of oxygen-blown autothermal reforming

at a pressure of about 40 atm, desulfurization and pre-reforming, a two-step Lurgi methanol synthesis in water- and gas-cooled reactors, and adjustment of syngas composition via hydrogen recycle. The reformer outlet temperatures range from about 950°C to 1,050°C. The synthesis gas is then compressed using a single casing gas compressor. An integrated recycler in the compressor generates the pressure for the methanol synthesis.

The methanol reactor contains a fixed number of tube sheets containing the copper-based catalyst. The isothermal reactor enables partial conversion of the syngas to methanol, reduces the by-products produced during the process, and increases the yields at low recycle ratios. The Lurgi-combined converter process employs a gas-cooled reactor and a water-cooled reactor. The gas containing methanol from the first reactor is transferred to a downstream water-cooled reactor where cooling takes place. The purge gas is separated and the cooled crude methanol is processed in a three-column distillation unit through which about 40% of the heating steam is saved.

As smaller, high-cost methanol plants continue to be replaced by mega-methanol plants, the industry's aggregate cash cost curve should drive down price to levels that would still provide adequate profitability at prices below \$100 per ton. This price environment creates new business opportunities for methanol to compete with other basic fuel and feedstock options. With combustion, a higher heating value of 9,776 Btu/lb, a \$100 per ton methanol fuel has an inherent energy cost of \$4.6 per million Btu. This cost is comparable to imported LNG at the receiving/regasification terminal. Since the transportation/distribution cost for methanol as a nonviscous liquid is relatively low, methanol fuel cost at the burner tip can be significantly cheaper than natural gas derived from LNG at the burner tip. Not only can methanol be considered under some economic scenarios as a primary fuel for combustion, its very low sulfur content (even when compared to ultralow sulfur diesel) provides an opportunity for methanol to be used as a backup fuel for utility and commercial power plants.

Research is also underway to convert synthesis gases to methanol in a liquid phase rather than using dry, fixed bed reactors. These technologies aim to avoid the



Gas to Liquid Technologies. Figure 4
The Lurgi MegaMethanol[®] process [41]

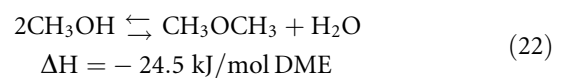
problem of low conversion rates and the need for recycling of gases.

Dimethylether

High oil prices and short supplies of LNG and LPG in Asia have led to the construction of many new DME plants, particularly in China. DME is currently industrially important as the starting material in production of the methylating agent dimethyl sulfate. DME has the potential to be used as a diesel or cooking fuel, a refrigerant, or a chemical feedstock. DME is also used as an aerosol propellant.

Reaction Pathway

DME traditionally is formed in a two-step process where methanol is first synthesized and is then dehydrated, over an acid catalyst such as γ -alumina, at methanol synthesis conditions. The dehydration reaction is shown in Eq. 22:

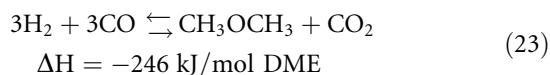


Alternatively, DME can be synthesized directly from syngas – without the separate process step

Gas to Liquid Technologies. Table 6 Comparative conversions for DME, methanol, and methanol/DME

Conversion	DME	Methanol	Methanol + DME
Per pass (%)	50	14	18
Total (%)	95	77	85

involving methanol production. Extensive studies have been performed to deduce the stoichiometry of direct DME synthesis from syngas, and the topic is still under debate [42]. Work by Air Products and JFE indicate that the stoichiometry is 1:1 in H_2/CO [43, 44]. The net reaction is shown in Eq. 23:



Net reaction 23 is a combination of Eqs. 20–22.

Because water produced in the dehydration step is consumed in the WGSR, the conversion of syngas can be higher in DME synthesis than in methanol synthesis. Table 6 shows the conversions obtained for direct processes producing DME, methanol, or a mixture of DME and methanol. The observed optimum ratio for DME synthesis is lower than that for methanol synthesis and ideally should be targeted at about 1.

Interestingly, the use of the methanol synthesis reaction 21 instead of 23 would give a net reaction for direct DME synthesis having H_2/CO stoichiometry of 2:1, which does not agree with experimental observation. Thus the direct synthesis of DME apparently proceeds through a different methanol-producing step than occurs in the production of methanol itself.

Commercial Activities in DME Synthesis

Gas phase reaction in a fixed bed was generally used in the past for DME synthesis from methanol. Development of the direct synthesis of DME in the gas phase with improved fixed bed reactors is also of interest [45]. Other improvements to direct DME production include the use of a slurry reactor for liquid phase, low-pressure DME synthesis.

Commercial production of DME originated as the formation of by-products during high-pressure methanol production. Direct, or single-step, synthesis of

DME from synthesis gas was developed in the 1980s, prompted by the high prices and uncertain supply of oil prevailing at the time. The objective of direct synthesis was to produce DME as an intermediate step en route to synthetic gasoline. Interest in non-oil-based fuels waned quickly as the oil supply position eased, until the early 1990s when interest was rekindled – this time on a much wider scale – by the increasing demand in Asia for LPG or the LPG-substitute, DME, and by the growing concern to produce cleaner fuels. The companies currently most actively engaged in the development of an economic DME process at large-scale are JFE (Japan), Haldor Topsøe (Denmark), and Air Products & Chemicals (USA).

Gas Phase DME Production Direct synthesis of DME in the gas phase with fixed bed reactors has been developed most extensively by Haldor Topsøe [45]. For large-scale processes, the reaction may be conducted in a series of three to four reactors with interstage cooling. Each reactor is operated adiabatically. Alternatively, the adiabatic operation may be carried out at low per pass conversion of CO, to keep the reaction exothermal within limits without interstage cooling (although such an option may not be economically attractive) [46].

Haldor Topsøe's process uses natural gas reforming for production of syngas with the ratio $(n_{H_2} - n_{CO_2})/(n_{CO} + n_{CO_2})$ between 2.04 and 2.1, where n is the number of moles. The greatest CO per pass conversion (65%) and the highest selectivity of DME relative to that of methanol (82/18) were observed using CO-rich syngas. With no CO_2 , the observed DME selectivity and productivity were substantially higher. Thus in contrast to the synthesis of methanol, described above, the presence of CO_2 in the feed syngas was found to have a major negative impact on the DME formation.

The current DME synthesis catalyst is formulated to catalyze the water gas shift reaction at a minimum level with minimal CO_2 production. The quality of the residual syngas from the reactors is therefore good enough for recycle without the need for a large, dedicated CO_2 removal plant. Ongoing improvements to the gas phase process include the development of bifunctional catalysts to produce DME in a single step in a single reactor [47].

Slurry Phase DME Production Air Products did a substantial body of research toward developing a slurry phase process for the direct synthesis of DME [48–52]. The presence of CO₂ in the feed gas was observed to have a major impact on DME formation, as expected based on the stoichiometry of Eq. 23. In the slurry reactor, syngas is bubbled upward through the solvent, which contains suspended catalyst particles. The catalyst is comprised of alumina particles of about 200 μm average size, with a layer of Zn/Cu oxide methanol synthesis catalyst formed around each alumina particle. The solvent is a light liquid oil (such as Witco-40) or a hydrocarbon such as *n*-hexadecane and acts as both a mass transfer medium for the reaction and a heat transfer medium through which heat from the exothermic reaction is removed by convection [53].

Control of the reaction temperature is even more important in DME synthesis than in methanol synthesis, because the higher equilibrium conversion to DME emits more heat and a hot spot in the reactor could damage the catalyst. In a slurry phase reactor, the heat of reaction is quickly absorbed by the high boiling liquid oil, which has a high heat capacity. Heat is removed from the reactor through immersed internal cooling coils, in which water is vaporized to steam. The temperature within the reactor vessel is well controlled in order to achieve higher conversion with longer catalyst life.

The direct synthesis technology under development by NKK Corporation since 1989 also uses a slurry phase reactor. In this process, natural gas is reformed in an autothermal reformer with oxygen and steam. Carbon dioxide is recycled back from the purification section of the plant to give synthesis gas with a H₂/CO ratio of 1.0. The standard DME synthesis reaction conditions are 260°C and 50 atm pressure. The condensed reactor product is purified in two distillation columns, with CO₂ vapor from the first column returned to the reactor, and the bottoms from the first column distilled in the second column. DME is condensed from the second column overheads, and the by-product methanol in the column bottoms returned to the DME synthesis reactor after water removal. The once through CO conversion is 50%. DME selectivity is more than 90%, and the molar ratio of DME to DME + methanol is 0.91. The coproduction of water is very small, with the

molar ratio of (DME/DME + methanol + water) = 0.013. With recycle the total CO conversion reaches 95%. A pilot demonstration plant project of 100 t/day DME capacity was begun by NKK at Hokkaido, Japan, in 2002 [54].

Mitsubishi Gas Chemical (MGC) touts higher single pass conversion for its Superconverter [50] slurry DME reactor system, claiming 70–80% conversion efficiency compared with 50–60% for other processes. MGC claims that DME produced at a remote site with low-cost natural gas feedstock in a 5,000–7,000 t/day plant should beat the delivered cost of both LNG and LPG in Japan. Japan DME (MGC, ITOCHU, and Mitsubishi Heavy Industries) have planned a 1 t/year capacity plant in Papua New Guinea.

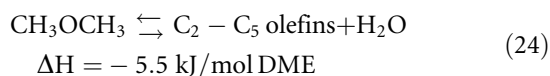
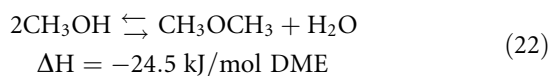
One key challenge in the economic design of 7,000 t/day plant is the reactor design configuration and catalyst to enable dehydration using a single reactor vessel, as opposed to multiple serial reactor stages or parallel reactor trains. Another important consideration is the ability to achieve sufficiently high methanol conversions to permit reaction product recovery in a single distillation train.

Another process for scales such as 7,000 t/day is referred to as the Jumbo process and is under development by Toyo Engineering using a dehydration reactor based on radial flow design. The process involves integrated coproduction of methanol from natural gas as an intermediate to DME. DME is produced via the vapor phase dehydration of methanol over a fixed bed of an acid catalyst, typically γ-alumina. Reaction temperatures are typically in the range of 250–400°C, with pressures as high as 250 psig. Gas space velocities have been reported to be in the range of 500–10,000 h^{−1} [56].

Toyo's DME process offers greater flexibility in the methanol/DME product distribution. Since the methanol and DME plants operate independently, the complex could be designed to produce methanol or DME exclusively or in any combination in accordance with market demand. Toyo's design approach would also permit the option of locating their Jumbo DME plant at a different location from that of the methanol plant. The cost of shipping methanol long distances may be less than the cost of shipping an equivalent quantity of DME due to the higher cost of the pressurized refrigerated storage requirements anticipated for oceangoing DME tankers.

Methanol to Gasoline and/or Diesel

The methanol to gasoline (MTG) process developed by Mobil Oil Corporation involves the conversion of methanol to hydrocarbons over a zeolite catalyst. The methanol conversion process occurs in two steps with chemistry as shown in Eqs. 22 and 24 (thermochemistry shown for ethylene as product).



Crude methanol (17% water) is superheated to 300°C and partially dehydrated over an alumina catalyst at 27 atm to yield an equilibrium mixture of methanol, diethyl ether, and water (75% conversion of methanol). This effluent is then mixed with heated recycled syngas and introduced into a reactor containing ZSM-5 catalyst at 350–366°C and 19–23 atm to produce hydrocarbons (44%) and water (56%). Because the zeolite has to be regenerated frequently to burn off coke formed during the reaction, the MTG process is rendered continuous by using multiple gasoline conversion reactors operating in parallel with 2–6 week cycles. Mobil's MTG process uses different pressures for syngas production (15–20 atm), methanol synthesis (50–100 atm), and the fixed bed MTG process step (15–25 atm).

The selectivity of Mobil's MTG process to gasoline range hydrocarbons is about 85+%, with the remainder of the product being primarily LPG. Nearly 40% of the gasoline are aromatic hydrocarbons with product distribution of 4% benzene, 26% toluene, 2% ethylbenzene, 43% xylenes, 14% trimethylbenzenes, plus 12% other aromatics. The shape selectivity of the zeolite catalyst results in a relatively high concentration of durene (1,2,4,5-tetramethylbenzene), representing about 3–5% of the gasoline produced. Therefore, MTG gasoline is usually distilled and the heavy fraction is hydroprocessed to reduce the concentration of durene to below 2%. This results in a high quality gasoline with a high-octane number. However, the future of the MTG process remains uncertain because the 1990 Clean Air Act Amendment limits the amount of aromatics in reformulated gasoline.

The first commercial MTG plant came onstream in 1985 in New Zealand (Mobil's Motunui plant), producing 4,500 t/day of methanol and 14,500 bbl/day of high-octane gasoline from natural gas. Gasoline production was later ceased and presently this plant and the nearby methanol plant at Waitara produce 2.43 million tons per year of chemical grade methanol for export [57].

A 100 bbl/day fluidized bed MTG pilot plant was jointly designed and operated near Cologne, Germany, by Mobil (supplying process technology and proprietary catalyst), Union Reinische Brankohlen Kraftstoff AG (acting as the operating agent), and Uhde GmbH (providing engineering services) from 1982 to 1985. Although no commercial plants have been built based on the fluid bed technology, it is considered ready for licensing and/or commercialization [58].

Mobil also developed a methanol to gasoline and diesel process, referred to as the MOGD process. Oligomerization, disproportionation, and aromatization of olefins are the basis for the MOGD process. Selectivity to gasoline and distillate from olefins is greater than 95%. The product slate from MOGD is 3 weight paraffins, 94 wt% olefins, 1 wt% naphthenes, and 2 wt% aromatics [59]. A large-scale test run was performed at a Mobil refinery in 1981. The MOGD process is not currently in commercial practice.

Haldor Topsøe's Integrated Gasoline Synthesis (TIGAS) process was developed to minimize capital and energy costs by integrating methanol synthesis with the MTG step into a single loop (i.e., without isolation of methanol as an intermediate). Whereas Mobil's MTG process uses different pressures for syngas production, methanol synthesis, and the fixed bed MTG process step, the TIGAS process involves catalysts and conditions modified such that the system pressure levels out and separate compression steps are not required. A mixture of methanol and DME is made prior to gasoline synthesis, which results in only one recycle loop from the gasoline synthesis step back to the methanol/DME synthesis step. A 1 t/day demonstration plant was built in Houston, Texas in 1984 and operated for 3 years. The gasoline yield for the TIGAS process (defined as the amount of gasoline produced divided by the amount of natural gas used for feed and as fuel) was 56%. The TIGAS process yields a lower

quality gasoline with a lower octane number compared to MTG, because of a reduced selectivity to gasoline range aromatics.

Methanol to Chemicals

The highest volume chemical processes using methanol as feedstock are formaldehyde, methyl tertiary-butyl ether, and acetic acid. Globally, formaldehyde production is the largest consumer of methanol, followed by methyl tertiary-butyl ether (MTBE) and acetic acid.

Formaldehyde is produced commercially from methanol by three industrial processes:

1. Partial oxidation and dehydrogenation with air in the presence of silver crystals, steam, and excess methanol at a temperature of 680–720°C, otherwise known as the BASF process. The methanol conversion for this process is 97–98%.
2. The same as process 1 except in the presence of crystalline silver or silver gauze at a temperature of 600–650°C. Then, the product is distilled and the unreacted methanol is recycled. The primary methanol conversion for this process is 77–87%.
3. Oxidation with only excess air in the presence of a modified iron/molybdenum/vanadium oxide catalyst at a temperature of 250–400°C, also known as the Formox® process (Perstorp Specialty Chemicals, Sweden). The methanol conversion for this process is 98–99%.

Formaldehyde is used to make resins with phenol, urea, or melamine for the manufacture of various construction board products.

MTBE is produced by reacting isobutene with methanol in the presence of an acidic catalyst. The reaction temperatures and pressures are 30–100°C, and 7–14 atm, respectively, so that the reaction occurs in the liquid phase. Catalysts used are solid acids, including zeolites (H-ZSM-5), and microporous sulfonic acid ion exchange resins such as Amberlyst-15. A molar excess of methanol is used to increase isobutene conversion and inhibit the dimerization and oligomerization of isobutene. At optimum reaction conditions, MTBE yields approaching 90% can be achieved. Currently there are over 140 MTBE plants, with a total installed capacity of about 20 million tons per year, using processes developed by Snamprogetti,

Huls (now Oxeno) ARCO, IFP, CDTECH (ABB Lummus Crest and Chemical Research Licensing), DEA (Formerly Duetsche Texaco), Shell (Netherlands), Phillips Petroleum, and Sumitomo.

Greater than 95% of the MTBE produced is used as a fuel additive in the gasoline motor pool. MTBE is also used in the petrochemical industry for the production of isobutene, and it also can be used in a number of chemical reactions including methacrolein, methacrylic acid, or isoprene production.

Acetic acid is produced by methanol carbonylation in the liquid phase using Co, Rh, or Ni catalysts promoted with iodine. The BASF process is initiated by the reaction of methanol with HI to yield methyl iodide. The active catalyst is the metal carbonyl $[\text{Rh}_2\text{I}_2(\text{CO})_2]$ into which methyl iodide inserts during the rate-limiting step. Acetic acid is formed by the hydrolysis of the eliminated acetyl iodide species CH_3COI that also regenerates HI. The process is run with over 99% selectivity at conditions of 180°C and 30–40 atm.

The Monsanto process is similar to that of BASF but is less severe. All new installed capacity since 1973 is based on the Monsanto process. Even so, the Rh/I catalytic system is very corrosive and requires expensive steels for materials of construction. Complete recovery of the expensive Rh catalyst and recycle of HI are paramount to maintain favorable process economics. The high cost of Rh has led to the search of other, lower cost metals that could be used as acetic acid process catalysts with similar performance.

Approximately half of the world's production of acetic acid comes from methanol carbonylation and about one third from acetaldehyde oxidation. Methanol carbonylation is the most likely technology of the future. Vinyl acetate, acetic anhydride, and terephthalic acid are all made from acetic acid. Latex emulsion resins for paints, adhesives, paper coatings, and textile finishing agents are made from vinyl acetate. Acetic anhydride is used in making cellulose acetate fibers and cellulosic plastics.

Ammonia

Ammonia is the basic building block of the nitrogen industry worldwide. Almost all ammonia is produced in the anhydrous form (free of water) by a catalytic reaction of nitrogen (from air) with hydrogen from

a hydrocarbon source (usually natural gas). Ammonia is a colorless, pungent, nonflammable gas at normal pressure and temperature and is lighter than air. For storage at atmospheric pressure at sea level, ammonia must be cooled to -33°C and stored as liquid. Lower temperatures are required at higher altitudes. In the usual atmospheric temperature range of 0 – 40°C , the range of vapor pressure is about 4–15 atm. Approximately 10% of the ammonia produced never reaches the market, because of its volatility and losses, during conversion to other materials and during transportation and storage. Nitrogen fertilizer consumption accounts for more than 85% of the world ammonia market.

General Process Description

The three main steps in producing ammonia are syngas production, syngas purification, and ammonia synthesis. The traditional and current commercial technology for hydrogen production in ammonia production involves steam reforming of natural gas to produce syngas (Eq. 2), and the water gas shift reaction (Eq. 19), both of which occur in the reformer. Steam reforming is by far the least expensive and most popular method of producing hydrogen for ammonia synthesis. Two-step steam reforming or partial oxidation also can be used to generate hydrogen-rich syngas for ammonia production, but these methods are most useful when the feed contains heavier hydrocarbons in addition to methane.

Raw synthesis gas must be purified before being sent to the ammonia synthesis unit. A shift conversion step is used to remove most of the carbon monoxide from the synthesis gas, because CO acts as a poison to the catalyst used in ammonia synthesis. Shift conversion by the WGSR (Eq. 19) also produces more hydrogen. The shift conversion is performed in two stages. In the first stage, called the high-temperature shift (300 – 450°C), the bulk of the carbon monoxide is oxidized to CO_2 over a chromium-promoted iron oxide catalyst. The second stage is a low-temperature shift (180 – 270°C) in which the remaining carbon monoxide is oxidized to CO_2 over a copper/zinc catalyst, leaving a few tenths of a percent in the syngas. Typical catalyst lifetimes for both high-temperature and low-temperature shift catalysts are 3–5 years.

Sulfur tolerant (“dirty shift”) catalysts also have been developed. These catalysts can handle larger amounts of sulfur. The company ICI makes dirty shift conversion catalysts that consist of cobalt and molybdenum oxides. They operate over a temperature range of 230 – 500°C . The controlling factors are the ratio of steam to sulfur in the feed gas and the catalyst temperature.

The synthesis gas is then washed with a solvent – such as monoethanolamine or other amine, potassium carbonate, or sulfolane (sulfolane and an alkanolamine) – to remove most of the carbon dioxide. The solvents typically contain an activator to promote mass transfer. Water scrubbing at high pressures was employed in the past to remove CO_2 but is no longer used. Most newer steam reforming plants use methanation to remove residual or trace carbon monoxide and carbon dioxide, by reaction with hydrogen over a nickel catalyst to yield methane and water.

Scrubbing with a liquid nitrogen wash produces a synthesis gas with very low inert gas content. This additional nitrogen is supplied by using air, in excess of the stoichiometric nitrogen requirement, in the secondary reformer or, in the case of plants with no reforming sections such as off-gas plants, by obtaining nitrogen from an associated air separation unit. Older processes involved scrubbing the synthesis gas with copper liquors, such as cuprous ammonium formate, to remove carbon monoxide, and caustic to remove carbon dioxide. Many modern processes include molecular sieve dryers downstream of the methanation step to remove the final traces of water. The synthesis gas is then compressed and finally reaches the ammonia converter, where the contained hydrogen and nitrogen chemically combine over a catalyst into ammonia.

Reaction Pathway

The chemical reaction considered in ammonia synthesis is shown in Eq. 25.

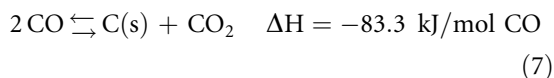


Since the reaction is exothermic, maximum conversion at equilibrium occurs at high pressure and low

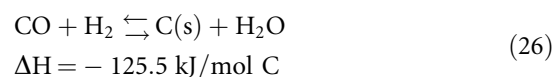
temperature. Ammonia synthesis is achieved in catalytic reactors at pressures ranging from 150 to 345 atm, at a minimum temperature of 430–489°C. A maximum temperature of ~500°C results from the balance between increasing reaction kinetics and decreasing the equilibrium ammonia concentration as the temperature increases. Thermodynamics suggest that a low process temperature is favored. However, the kinetics of the reactions dictates that high temperature is required. Nitrogen is usually fed as air, but in some plants (e.g., partial oxidation plants) the N₂ is a purified gas from an associated air separation facility. In terms of the hydrogen source, production methods, product recovery, and overall plant engineering, there are many process variants for ammonia production.

About 50% of the hydrogen in ammonia comes from steam. High temperatures and low pressures favor the highly endothermic methane reforming reaction. Higher pressures tend to lower the methane conversion. In industrial reformers, the reforming and shift reactions result in a product gas composition that closely approaches equilibrium. However, the following side reactions produce carbon in the steam reformer:

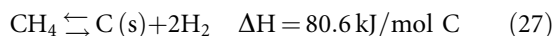
Boudouard coking



CO reduction



Methane cracking



The molar steam to carbon ratio in the reformer is usually between 2 and 6, depending on the feedstock and process operating conditions. Excess steam is used to prevent coking in the reformer tubes. The shift reaction is exothermic and favors low temperatures. Since the shift reaction does not approach completion in the reformer (usually there is 10–15 vol% CO, dry basis, in the reformer effluent), further conversion of CO is performed using external shift conversion catalysts.

Reactors and Catalysis for Ammonia Synthesis

The basic design of an ammonia synthesis reactor is a pressure vessel with sections for catalyst beds and heat exchangers. Over the years, there have been many different designs proposed and constructed for commercial operation such that today, ammonia reactors are classified by flow type (radial, axial, or cross flow) and cooling method (quench or indirect) used.

Axial flow reactors are essentially top-to-bottom flow reactors. The design is comparatively simple; however, a fairly large pressure drop develops across the catalyst bed. A radial flow configuration feeds gas into an annular region between the reactor wall and the outer surface of the catalyst bed. Gas flows through the bed and exits out a central collection tube. This design minimizes pressure drop across a shallow bed with a large surface area. Radial flow reactors tend to be tall vessels with relatively small diameters. The cross-flow reactor configuration has a similar principle in that gas is introduced along one side of the reactor and is collected radially across the reactor by a collector on the other side.

Reactant gas is preheated either by circulating it through heat exchangers or by using water to produce steam. However, removing the heat generated from the exothermic synthesis reaction to maintain control of the reaction temperature is a challenge. Quench converters are designed to introduce cool reactant gas at various points along the length of the catalyst bed. Interbed heat exchangers also can be used to remove heat at specific intervals along the bed, effectively separating the bed into multiple synthesis zones, or continuously along the bed with cooling tubes. These indirectly cooled designs allow for efficient recovery of reaction heat that can be used in other parts of the process.

Early designs based on axial flow were easier to build but not as efficient as the more complex radial flow reactors that have more recently been designed. One example of a radial flow design, by Haldor Topsøe, uses two radial beds with quench gas injection between them. A similar radial flow design uses an interbed heat exchanger in the first catalyst bed. Cold ammonia synthesis gas is introduced from the bottom of the reactor through the second catalyst bed, then through the heat exchanger in the first catalyst bed. The cold gas flow

through the second bed also provides indirect heat exchange. An additional heat exchanger is located at the bottom of the reactor to cool the reactor gases.

The Kellogg four-bed axial flow reactor system is a good example of an axial flow reactor design. This quench reactor consists of four catalyst beds held on separate grids. Quench gas is introduced in the spaces between the bed, and a heat exchanger is located at the top of the vessel. Another type of Kellogg reactor is a cross-flow reactor design where gas flows through the catalyst bed perpendicular to the vessel axis. It is available in both quench and indirectly cooled versions.

The ammonia synthesis catalyst is commonly iron, such as magnetite, which may be promoted with aluminum, potassium, and/or calcium. Finely divided iron is pyrophoric, therefore Fe catalysts are reduced to metallic form in situ. Porosity is developed as the catalysts are reduced. During reduction, oxygen is removed from the magnetic crystal lattice without shrinkages. Metallic Fe is formed with essentially the same porous structure as the magnetite precursor. The production and preservation of this highly porous structure during reduction of the ammonia synthesis catalyst precursor leads to highly active catalysts. The addition of structural promoters facilitates the formation of highly porous metallic iron.

Catalyst activity is not directly associated with the Fe surface area but is related to complex interactions of the promoters. Alkali metal promoters in ammonia synthesis catalysts are necessary to attain high activity. Potassium is the most effective alkali promoter, while Li and Na are poor promoters and are not used commercially. Potassium is thought to interact with the Fe crystallites to increase the dissociative sticking probability of N_2 on the Fe sites and thereby increase catalytic activity [60].

The most effective catalysts from a process perspective are those that have the highest rate of conversion at the lowest temperatures. Other catalysts are being developed such as promoted Ru on high surface area graphite supports, and these recently commercialized catalysts offer the possibility of greatly decreasing synthesis temperatures and pressures with improved activity at high ammonia concentrations. The Kellogg Advanced Ammonia Process is based on a Ru catalyst that is claimed to be 40%

more effective than Fe catalysts, because synthesis efficiency is increased by lowering the process pressure from 150 to 90 atm.

With effective gas cleanup and conditioning, commercial ammonia synthesis catalyst lifetimes of 5–8 years can be achieved in most cases and potentially up to 14 years. Ammonia formation increases as the process pressure is increased. The optimum H_2/N_2 ratio is near 2 at high-space velocities and approaches 3 at low-space velocities as equilibrium becomes more dominant. Space velocities in commercial ammonia synthesis vary from 12,000/h at 150 atm to 35,000/h at 790 atm.

Per pass conversion on the order of 10–35% is typically achieved. Ammonia is recovered from the synthesis loop by cooling to condense the synthesis gas at process pressures. The liquid ammonia is separated from the gas effluent, which is recycled back through to the reactor. Earlier plant designs used air or water cooling for ammonia recovery. Modern synthesis plants use refrigeration to condense out the ammonia. The ammonia recovery is not extremely efficient so the recycled gas typically contains 4% ammonia plus any inert gasses (Ar, He, CH_4 , etc.) that may be in the process stream. Purging some of the gas in the recycle loop before it is recycled minimizes inert gas concentrations, and flashing the liquid ammonia in a pressure letdown step releases any dissolved gases.

Clearly, the ammonia synthesis process consists of many complex unit operations apart from the actual synthesis loop. The way in which these process components are combined with respect to mass and energy flow has a major influence on efficiency and reliability. Many of the differences between various commercial ammonia processes lie in the way in which the above process elements are integrated.

Commercial Activities in Ammonia Synthesis

The entire ammonia industry was originally based on using coal as a feedstock. The industry has now moved toward natural gas as the main hydrocarbon feedstock. By 1990, only 14% of the world ammonia capacity was based on coal or coke. Apart from a few plants operating in India and South Africa, today the majority of coal-based ammonia plants are found in China.

More than 20 commercial ammonia synthesis processes have been described [61].

Commercial technology vendors for ammonia include:

- Johnson Matthey
- Linde
- Kellogg Brown and Root
- Haldor Topsøe
- Ammonia Casale
- Uhde

As a result of a surge of major developments in ammonia process technology beginning in the late 1960s, large 1,000–1,500 t/day capacities became the industry standard for new plant constructions. Plants as large as 2,000 t/day are now common. The main ammonia process contractors have each now developed “mega-ammonia technology,” and are offering large plant concepts from 3,000 to 6,000 t/day units. The world’s largest ammonia plant, in Al Jubail, Saudi Arabia, is a collaboration between Saudi Basic Industries Corporation (SABIC) and Saudi Arabian Fertilizer Company. The plant, called SAFCO IV, came onstream in 2007 and has a capacity of 3,300 t/day. This huge plant is the first application of a new ammonia process from Uhde, referred to as the Dual Pressure Process, and utilizes Johnson Matthey’s KATALCO™ catalysts [62]. A major factor resulting in much lower average production costs has been the switch from electrically driven compression to the use of steam-driven centrifugal compressors powered by waste heat. During the past 30 years, many of the smaller, older, higher-cost plants in the world have been shut down.

Developments to improve energy efficiency or overcome interruptions of natural gas supply have been examined and some are finding application. The recovery of the hydrogen and ammonia from the synthesis purge gas by a cryogenic unit or membrane is now standard and an additional 5% of ammonia can be produced from the same amount of feedstock. The use of computer controls to improve overall plant performance is increasing.

Substantial improvements have been made over the years in the energy efficiency of carbon dioxide removal systems. The first large-scale ammonia plants in the 1960s typically used monoethanolamine as a solvent, with energy input of over 50,000 kcal/kg-mol of carbon

dioxide removed, whereas today’s modern plants use improvements such as the dual-step methyl diethanolamine process that can reduce the energy input to about 10,000 kcal/kg-mol or lower of carbon dioxide removed. As described above, improvements to efficiency also have been introduced with new ammonia synthesis catalysts.

The current generation of large ammonia plants is much more fuel-efficient than plants constructed 20 or 30 years ago. A typical world-scale plant constructed in the 1970s consumed about 42 million Btu of natural gas per ton of ammonia produced. Retrofitting such a plant to improve fuel efficiency can reduce gas consumption to about 36 million Btu/t; however, the newest generation of ammonia plants use only about 30 million Btu/t of ammonia, are reported to be easier to operate, and have slightly lower conversion costs. Some newer plants also recover more than 1 million Btu/t by generating electricity, used elsewhere in the complex, from waste heat.

An additional minor trend has been to construct an ammonia plant at a location where a hydrogen off-gas stream is available from a nearby methanol or ethylene operation (e.g., Canadian plants at Kitimat, British Columbia, and Joffre, Alberta). Gas consumption at such production sites ranges from 25 to 27 million Btu/t of ammonia, depending on specific circumstances. Perhaps more important, the capital cost of such a plant is only about 50% of the cost of a conventional plant of similar capacity because only the synthesis portion of the plant is required.

Direct Conversion of Natural Gas to Liquids

Substantial capital cost could be avoided if methane could be directly converted to methanol, or other high-volume liquid fuels or chemicals, without the intermediate production of syngas. Several direct processes have been investigated over the years, including methane coupling to produce ethane, which can be oligomerized to give gasoline in a reactor similar to that used in the MTG process. Another consideration has been the direct production of aromatics, using an oxidative coupling membrane. Another option is the selective oxidation of natural gas to produce fuels directly. Direct methods

would need to have a distinct economic advantage over indirect methods, but to date no direct processes have progressed to a commercial stage. Product yields are generally small while operating in the single pass mode, which makes separations difficult and costly.

Oxidative Coupling to Light Hydrocarbons

In the oxidative coupling reaction, methane and oxygen react over a catalyst at elevated temperatures (700°C) and normal pressures to form ethane as a primary product and ethylene as a secondary product. British Petroleum (now BP) has described experiments at higher temperatures (1,100°C) and short residence times, at which conditions methane reacts with oxygen to produce syngas as well as C_{2+} hydrocarbons in addition to syngas. Several catalysts including zirconia gave C_{2+} selectivities of over 50%.

Work on the lower temperature methane coupling reaction has been done by Phillips Petroleum, Texas A&M University, Akzo Chemie, Amoco (now BP), the University of California at Berkeley, the University of Pittsburg, the University of Tokyo, Idmitsu Kosan, and Union Carbide among others. Unfortunately both the methane and the ethylene may react further to produce CO_2 , and the single pass combined yield of ethylene and ethane is limited to about 25%. With better catalysts, such as SrO/La_2O_3 and $Mn/Na_2WO_4/SiO_2$, a C_2 selectivity of about 80% can be achieved, but at a low methane conversion of 20%. About half of the C_2 product is ethane and half ethylene. The best C_2 selectivities are almost always achieved under oxygen-limiting conditions. Because the reaction is exothermic, a zone within the catalyst bed may be 150–300°C hotter than the external temperature and heat management is a serious problem. This is complicated by the fact that metals normally used for the construction of reactors catalyze the combustion of methane.

The yield of ethylene can be improved considerably by operating in a recycle mode with continuous removal of ethylene, and by applying various techniques for separating products from reactants including cryogenic, membrane, and adsorptive separations. For economic reasons, either molecular sieve or membrane separation systems are most likely to be

employed, both of which have severe mass transport limitations not suitable for scale-up. A variation of this recycle system is to use a zeolite catalytic reactor to convert ethylene to benzene and toluene. In principle, aromatic yields in excess of 70% could be achieved; however, to achieve these yields the conversion of ethane to ethylene must be highly selective.

Selective Oxidation of Natural Gas to Methanol

Selective alkane oxidation by transition metal complexes in solution has been the focus of substantial effort since the 1970s. However, the low activity of alkanes and the typically higher activity of the desired products make this process a great challenge. In 1998, Periana et al. of Catalytica Inc. reported developing an effective catalyst for selective oxidation of CH_4 to CH_3OH in high yield. The initial catalyst was dichloro- $(\eta^2-(2,2'-bipyrimidyl))$ platinum(II) complex, $(bpy)mPtCl_2$. In well-dried sulfuric acid (102%), CH_4 at 3,400 kPa and 220°C was converted to a mixture of CH_3OSO_3H (methyl bisulfate, which presumably would be subsequently hydrolyzed to methanol) and CH_3OH at 72% conversion with selectivity of 81% [63]. Although catalytic in Pt complex, the reaction was very slow and required over half as many moles of catalyst as moles of methane converted. Another catalyst, $(NH_3)_2PtCl_2$, system was more active but less stable. A complete cycle would require the regeneration of concentrated sulfuric acid. The Catalytica process is promising, because it potentially has high yield at relatively low temperatures. On the other hand, it suffers from drawbacks including (1) the catalytic rate is too low; (2) the separation cost is too high; (3) the solvent is too corrosive; (4) reoxidation of SO_2 with regeneration of concentrated sulfuric acid is too expensive; and (5) the catalyst is severely inhibited by water.

Cold flame oxidation is another approach to direct conversion of methane to methanol. A pressurized mixture of methane and oxygen is reacted at moderate temperatures of 350–500°C. The reaction mixture is very fuel rich, with methane to oxygen ratios of about 20:1. The main chemical reaction is the oxidation of methane to methanol. However, further oxidation of methanol to formaldehyde often takes place simultaneously. The University of Manitoba reported 90% selectivity for methanol at a single pass conversion of 7.5% in an isothermal

process. Although this is among the best results reported to date, the yield of methanol from this process is too low to be of commercial interest.

Future Directions

The immediate future of F-T and DME gas to liquids would seem to be in their implementation. Acquisitions and ventures are creating “one stop shops” for commercialization. Johnson Matthey has acquired Davy Process technology and ICI, making the parent company a leader in reforming and methanol GTL technologies. The companies Enichem and IFP/Axens have strong capabilities particularly applicable to F-T GTL. Enichem has skills in catalysis, engineering, and the oil and gas business. IFP/Axens has catalysis expertise, plus basic engineering, process development, licensing, and catalyst manufacturing [64]. The future of methanol may be the status quo, because the technology is already mature. Alternatively, GTL methanol plants may expand in number (and possibly size) given the anticipated increased availability of unconventional gas combined with high oil process.

Bibliography

- Crane H, Kinderman E, Malhotra R (2010) A cubic mile of oil. Oxford University Press, Oxford
- United States Environmental Protection Agency (1998) Compilation of emission factors AP-42, v1, 5th edn, Supplement D
- United States Environmental Protection Agency (2010) Compilation of emission factors AP-42, v1, 5th edn, Supplement E, corrected. Calculation for low-sulfur No. 6 fuel oil
- United States Environmental Protection Agency (1998) Compilation of emission factors AP-42, v1, 5th ed, Supplement E. Calculation for medium volatile bituminous coal
- Rajnauth J, Ayeni K, Barrufet M (2008) Gas transportation: present and future. In: CIPC/SPE Gas Technology Symposium 2008 Joint Conference, Calgary, Alberta, 16–19 June 2008
- Bellussi G, Zennaro R (2007) New developments: energy, transport, sustainability. In: Encyclopaedia of Hydrocarbons, vol III, Chap. 2.6, pp. 161–182, EniChem
- Smith R, Asaro M (2005) Fuels of the future: technology intelligence for gas to liquids strategies. SRI Consulting, Menlo Park
- Yost C, DiNapoli R (2003) Benchmarking study compares LNG plant costs. Oil Gas J 101(15):56–59
- Nielsen R (2001) Fundamentals of mixed refrigerant compared to conventional refrigeration are discussed in “Ethylene Plant Enhancement.” PEP Report 29 G, SRI Consulting, Menlo Park
- Low WR, Andress D, Houser C (1997) Method of load distribution in a cascaded refrigeration process. US 5611216 to Phillips Petroleum Company, 18 Mar 1997
- Houser C, Yao J, Andress D, Low WR (1997) Efficiency improvement of open-cycle cascaded refrigeration process. US 5669234 to Phillips Petroleum Company, 23 Sept 1997
- Delong BW (1987) Method for cooling normally gaseous material. US 4680041 to Phillips Petroleum Company, 14 July 1987
- Netzer D, Nielsen R (2003) Baseload liquefied natural gas by cascade refrigeration. PEP Review 2003–15, SRI Consulting, Menlo Park
- Smith R, Asaro M (2005) Fuels of the future: technology intelligence for gas to liquids strategies. SRI Consulting, Menlo Park
- Huffman GP, Feeley III TJ (2000) Fuel science in the Year 2000: where do we stand, where do we go from here? I: Power generation and related environmental concerns – DOE’s fine particulate and air toxics research program: responding to the environmental challenges to coal-based power production in the 21st century. Preprints of the Division of Fuel Chemistry of the American Chemical Society, 45(1):108–112, 2000 Spring conference of the American Chemical Society, San Francisco
- Topsoe_synthesis_g#6D6FFA1.ashx.pdf. Accessed 2 Apr 2011, reprinted from Hydrocarbon Engineering, 2006
- Christensen TS, Østberg M, Bak Hansen J-H (2001) Process demonstration of autothermal reforming at low steam-to-carbon ratios for production of synthesis gas. In: AIChE Annual Meeting, Reno, 4–9 Nov 2001
- Wesenberg MH (2006) Gas heated steam reformer modelling. PhD thesis, The Norwegian University of Science and Technology
- Aasberg-Petersen K, Christensen TS, Charlotte Stub Nielsen CS, Dybkjær I (2003) Recent developments in autothermal reforming and pre-reforming for synthesis gas production in GTL applications. Fuel Process Technol 83(1–3):253–261
- Loock S, Ernst WS, Thomsen SG, Jensen MF (2005) Improving carbon efficiency in an auto-thermal methane reforming plant with gas heated heat exchange reforming technology. Paper No. O96-001, 7th World Congress of Chemical Engineering, Glasgow
- Tsuru T, Yamaguchi K, Yoshioka T, Asaeda M (2004) Methane steam reforming by microporous catalytic membrane reactors. AIChE J 50(11):2794–2805
- Carolan MF, Chen CM, Rynders SW (2003) Development of the ceramic membrane ITM syngas/ITM hydrogen process. Fuel Chem Div Preprints 48(1):344
- Robinson ET (S), Sirman J, Apte P, Gui X, Bulicz TR, Corgard D, Hemmings J (2005) Development of OTM syngas process and testing of syngas derived ultra-clean fuels in diesel engines and fuel cells. DE-FC26-01NT41096, Final Report
- Caro J, Wang H, Noack M, Koelsch P, Kapteijn F, Kannelopolous N, Nolan J (2007) Manufacture of composite membranes and their use for selective partial oxidation reactions of hydrocarbons. EP 1847311 to Universität Hannover, Germany
- Dupont V, Ross AB, Knight E, Hanky I, Twigg MV (2008) Production of hydrogen by unmixed steam reforming of methane. Chem Eng Sci 63(11):2966–2979
- <http://www.statoil.com/en/OurOperations/TerminalsRefining/Tjeldbergodden/Pages/default.aspx>. Accessed 11 Jan 2011

27. Davis BH (2001) Fischer–Tropsch synthesis: current mechanism and futuristic needs. *Fuel Process Technol* 71:157–166
28. Brady RC III, Pettit R (1981) The chain propagation step. *J Amer Chem Soc* 1981:287–1289
29. Davis BH (2001) Fischer–Tropsch synthesis: current mechanism and futuristic needs. *Fuel Process Technol* 71:157–166
30. Oukaci R, Singleton AH, Goodwin JG Jr (1999) Comparison of patented Co F–T catalysts using fixed-bed and slurry bubble column reactors. *Appl Catal A General* 186:129–144
31. Manzer L, Schwarz, S (2002) Fischer–Tropsch processes using catalysts on mesoporous supports. US 2002052289 to Conoco, 2 May 2002
32. <http://www.zawya.com/projects/project.cfm/pid210307061231?cc>. Accessed 11 Jan 2011
33. Spath PL, Dayton DC (2003) Preliminary screening – technical and economic assessment of synthesis gas to fuels and chemicals with emphasis on the potential for biomass-derived syngas. NREL/TP-510-34929
34. Fisher IA, Bell AT (1998) In situ infrared study of methanol synthesis from H_2/CO over Cu/SiO_2 and $Cu/ZrO_2/SiO_2$. *J Catal* 178(1):153–173
35. Chinchin GC, Denny PJ, Parker DG, Spencer MS, Whan DA (1987) Mechanism of methanol synthesis from $CO_2/CO/H_2$ mixtures over copper/zinc oxide/alumina catalysts: use of ^{14}C -labelled reactants. *Appl Catal* 30(2):333–338
36. Grabow LC, Mavrikakis M (2011) Mechanism of methanol synthesis on Cu through CO_2 and CO hydrogenation. *ACS Catal* 1(4):365–384
37. Tijm PJA, Waller FJ, Brown DM (2001) Methanol technology developments for the new millennium. *Appl Catal A General* 221:275–282
38. Liu J, Wei R, Zhang Y, Xu R, Li Z (2009) Preparation of $Cu/ZnO/Al_2O_3$ catalysts for methanol synthesis by improved two-step coprecipitation method. *Gongye Cuihua* 17(7):22–25. C.A. 2009:1625358
39. Baltes C, Vukojevic S, Schüth F (2008) Correlations between synthesis, precursor, and catalyst structure and activity of a large set of $CuO/ZnO/Al_2O_3$ catalysts for methanol synthesis. *J Catal* 258(2):334–344
40. Kaluza S, Behrens M, Schiefenhoevel N, Knip B, Fischer R, Schloegl R, Muhler M (2011) A novel synthesis route for $Cu/ZnO/Al_2O_3$ catalysts used in methanol synthesis: combining continuous consecutive precipitation with continuous aging of the precipitate. *ChemCatChem* 3(1):189–199
41. Lurgi brochure 0312e_MegaMethanol.pdf
42. Smith R, Naqvi S, Asaro M (2008) Fuels of the future: technology intelligence for coal to liquids strategies. SRI Consulting, Menlo Park
43. Lewnard JJ, Hsuing TH, White JF, Brown DM (1990) Single-step synthesis of dimethyl ether in a slurry reactor. *Chem Eng Sci* 45(8):2753–2741
44. Ohno Y, Omiya M (2003) Coal conversion into dimethyl ether as an innovative clean fuel. In: 12th ICCS Coal Conversion in DME, 2–6 Nov 2003
45. Haugaard J, Voss B (2001) Process for the synthesis of a methanol/dimethyl ether mixture from synthesis gas. US 6191175 to Haldor Topsøe, 20 Feb 2001
46. Naqvi S (2002) Dimethyl ether as alternative. Fuel PEP Report 245, SRI Consulting, Menlo Park
47. Kang S-H, Bae JW, Kim H-S, Dhar GM, Jun K-W (2010) Enhanced catalytic performance for dimethyl ether synthesis from syngas with the addition of Zr or Ga on a $Cu - ZnO - Al_2O_3/\gamma-Al_2O_3$ bifunctional catalyst. *Energy Fuels* 24(2):804–810
48. Bhatt BL, Schaub E, Heydorn E (1993) Recent developments in slurry reactor technology at the LaPorte Alternative Fuels Development Unit. In: International Technical Conference on Coal Utilization & Fuel Systems, LaPorte, pp 197–208, 26–29 Apr 1993
49. Peng X-D, Toseland B, Underwood T (1997) A novel mechanism of catalyst deactivation in liquid phase synthesis gas-to-DME reactions. In: Bartholomew C, Fuentes GH (eds) Catalyst deactivation. Elsevier, Amsterdam
50. Peng X-D (2002) Catalyst activity maintenance for the liquid phase synthesis gas-to-dimethyl ether process. Part II: Development of aluminum phosphate as the dehydration catalyst for the single-step liquid phase syngas-to-DME process. DOE contract DE-FC22-94PC93052, Final Report
51. Peng X-D (2002) Kinetic understanding of the syngas-to-DME reaction system and its implications to process and economics. DOE Contract DE-FC22-94 PC93052, Topical Report
52. Tijm PJ (2003) Development of alternative fuels and chemicals from synthesis gas. DOE Contract number FC22-95PC93052, Final Report
53. Smith R (2009) Dimethyl ether (DME) from coal. PEP Report 245B, SRI Consulting, Menlo Park
54. Ogawa T, Inoue N, Shikada T, Ohno Y (2003) Direct dimethyl ether synthesis. *J Nat Gas Chem* 12:219–227
55. The Ministry of Industry, Energy and Tourism; Orkustofnun/The National Energy Authority, The Innovation Center Iceland; Mitsubishi Heavy Industries, Ltd.; Mitsubishi Corporation; Hekla hf.; NordicBlueEnergy (2010) A feasibility study report for a DME project in Iceland. IDME Project Feasibility Study – 2009. Accessed 4 Apr 2011
56. Pavone T (2003) Jumbo dimethyl ether production process via Toyo technology. PEP Review 2003–9, SRI Consulting, Menlo Park
57. http://www.methanex.com/ourcompany/locations_newzealand.html. Accessed 4 Apr 2011
58. More detailed process design and economics information can be found in Apanel G (1999) Liquid hydrocarbons from synthesis gas. PEP Report 191A, SRI Consulting, Menlo Park
59. Tabak SA, Yurchak S (1990) Conversion of methanol over ZSM-5 to fuels and chemicals. *Catal Today* 6(3):307–327
60. Spath P, Dayton D (2003) Bioproducts from syngas, Syngas_products.pdf. Accessed 2 Apr 2011
61. Ullmann's Encyclopedia of Industrial Chemistry (2002) "Ammonia" published online 15 Dec 2006, doi: 10.1002/14356007.a02_143.pub2. Accessed 4 Apr 2011

62. Shah J (2007) SAFCO IV: catalyst start-ups in the world's largest ammonia plant. In: 20th AFA International Annual Technical Conference, Tunisia. 5_03 John_BRIGHTLING_ Johnson Matthey Catalysts_ U.K.pdf. Accessed 4 Apr 2011
63. Xu X, Fu G, Goddard III WA, Periana RA (2004) "Selective oxidation of CH₄ to CH₃OH using the Catalytica (bpym)PtCl₂ catalyst: a theoretical study" in Studies in surface science and catalysis, Natural gas conversion VII. In: Proceedings of the 7th Natural Gas Conversion Symposium, Dalian, vol 147, pp 499–504
64. Zennaro R, Hugues F, Caprani E (2006) The Eni – IFP/Axens GTL technology: from R&D to a successful scale-up. In: DGMK – SCI conference on synthesis gas chemistry, Dresden

Gasification and Liquefaction Alternatives to Incineration in Japan

KUNIO YOSHIKAWA

Frontier Research Center, Tokyo Institute of Technology, Yokohama, Japan

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

EBARA Fluidized Bed Gasification and Ash-Melting Process

The JFE High Temperature Gasifying and Direct Melting Process

The TOSHIBA Process for Liquefaction of Plastic Wastes

Future Directions

Bibliography

Glossary

Gasification Thermal process that involves the reaction of carbonaceous feedstocks with oxygen-containing reagents, usually air, oxygen, steam, or carbon dioxide, generally at temperatures in excess of 800°C

MSW Municipal solid waste

PET Polyethylene terephthalate

PVC Polyvinylchloride

Pyrolysis Thermal process that implies the degradation of the organic materials at temperatures in the

range of 400–800°C and in the absence of oxygen or other reagents

RDF Refuse derived fuel

Slag Molten ash

SR Shedder residue

Definition of the Subject and Its Importance

The major technologies used in Japan for energy recovery from municipal solid waste (MSW) are mass burning incinerators combined with landfilling of ash. However, shortage of landfill space along with new regulations for dioxin emission control and the Japanese Containers and Packaging Recycling Law has stimulated active R&D and commercialization of relatively novel thermal treatment processes based on gasification and liquefaction of MSW.

The purpose of this article is to introduce novel gasification and liquefaction processes for MSW that are already commercialized in Japan and are potential future alternatives to mass burning for effective resource recovery from MSW.

Introduction

In Japan, about 40 million tons of municipal solid wastes (MSW) are incinerated each year. Of these, about 20 million tons are used for power generation and in total about 1,000 MW of electric power is produced from MSW. Most of these waste-to-energy (WTE) plants are large-scale plants exceeding 200 t/day scale.

The major technologies used in WTE plants in Japan are stoker-type mass burning of as received MSW, where the final residues such as ash are landfilled. However, shortage of landfill space and also new regulations for detoxifying the fly ash by-product of incineration, as of 2004, has driven many municipalities to accept relatively novel processes such as direct gasification and smelting and, also, rotary kiln or fluidized bed gasification combined with melting of the ash to a vitrified slag.

On the other hand, many municipalities have started to source-separate and collect the plastic materials contained in MSW, under the Japanese Containers and Packaging Recycling Law enacted in 1995. The segregated plastic materials are recycled by three methods: material recycling, chemical recycling, and

production of solid fuel. One of the technologies in the chemical recycling technology is liquefaction of plastic wastes.

There are about 30 Japanese companies engaged in the development of gasification and ash-melting systems, which can be divided into three main types: (1) the vertical shaft types that melt the entire amount of wastes directly, (2) the fluidized bed types that gasify the wastes directly with slagging, and (3) the kiln types that gasify the wastes indirectly with slagging.

This article introduces three novel MSW thermal treatment processes developed and commercialized in Japan: (1) The EBARA Fluidized Bed Gasification and Ash-melting Process, (2) The JFE High Temperature Gasifying and Direct Melting Process, and (3) the TOSHIBA Waste Plastics Liquefaction Process. These processes have the potential to be future alternatives to the existing mass burning processes for maximizing the effective recycling and utilization of MSW.

EBARA Fluidized Bed Gasification and Ash-Melting Process

Process Description

Since the year 2000, EBARA's Fluidized Bed Gasification and Ash-melting process (TwinRec process) is in operation in large commercial installations [1]. It is based on fluidized bed gasification in combination with ash melting. The following description is focused on the core components of the TwinRec system: the fluidized bed gasifier and the cyclonic ash-melting chamber.

Any type of waste can be fed into the gasifier. Only bulky wastes need to be cut to pieces smaller than 30 cm in length. The gasifier is a proprietary internally circulating fluidized bed of compact dimensions, operated at temperatures between 500°C and 600°C. The resulting syngas (fuel gas) and fine particles are entrained into the gas flow leaving the gasifier. The low gasification temperature in the fluidized bed leads to easily controllable process conditions.

The main function of the gasifier is to separate the combustible gases and the dust from the inert and metallic particles of the waste. Metals contained in the waste, such as aluminum, copper, and iron, can be recycled as valuable products from the bottom off-stream of the gasifier as they are neither oxidized

nor sintered with other ash components. Together with these metals, larger inert particles are removed from the furnace. Smaller inert particles are returned to the gasifier where they serve as bed material. The fine inerts are blown out of the gasifier and enter the next stage of the process.

Figures 1 and 2 show the operating principle of the gasifier and the ash-melting furnace and the process flowsheet, respectively. The fuel gas and carbonaceous particles that are produced in the gasifier are combusted in the cyclonic ash-melting chamber at temperatures between 1,350°C and 1,400°C by the addition of secondary air. Here, the fine particles are collected on the walls, where they are vitrified and slowly flow downward through the furnace.

The molten slag collected in the furnace is then quenched into a water bath to form a granulate with excellent leaching resistance; this vitrified material meets all regulations for recycling in construction.

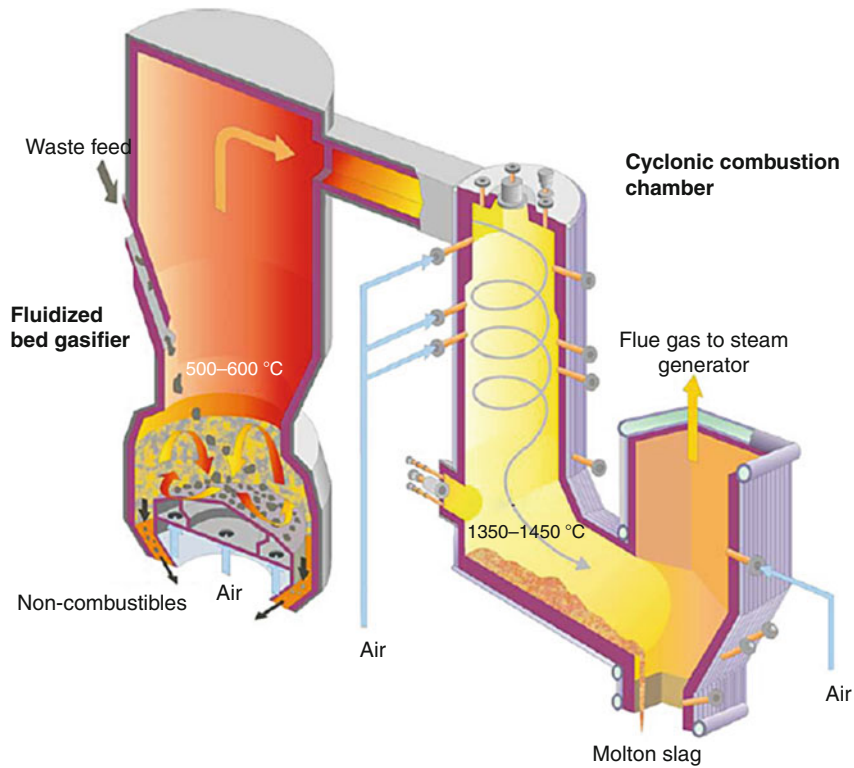
The high combustion temperature ensures that the most stringent dioxin emission regulations, below 0.1 ng TEQ/Nm³ are met by means of minimal air pollution control measures.

The gasifier and the ash-melting furnace operate at atmospheric conditions, without any auxiliary fuels, except for start-up of the process or industrial oxygen. Due to the low excess air ratio that is required for complete combustion, the steam generator, boiler and air pollution control system are very compact. The energy content of the waste is converted into electricity and/or district heat with a high net efficiency.

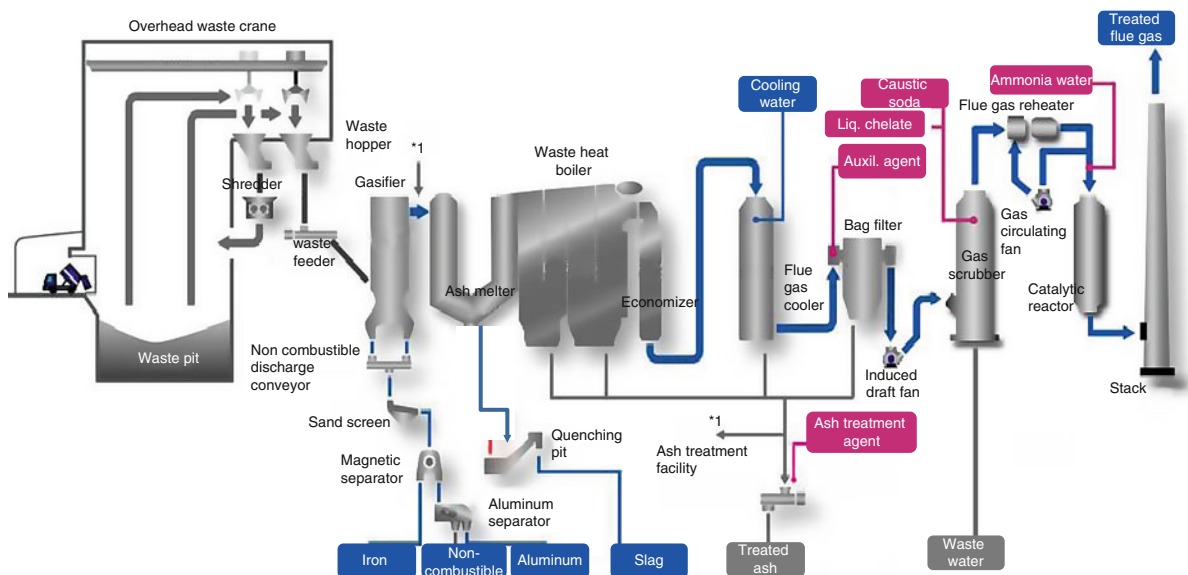
Recycling and Recovery

The Ebara TwinRec process can treat a wide range of materials generate product streams that match their characteristics and enable optimal resource recovery:

- Metals and alloys are not oxidized in the gasifier and can be recycled.
- Inert mineral materials are free of dust and organic matter and are also suitable for recycling.
- Mineral dust and metal oxide powder are vitrified into slag and can be used as construction materials.
- Any toxic organic substances are completely destroyed and the total organic content is transformed into energy.



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 1
The fluidized bed gasifier and the ash melting furnace of EBARA TwinRec process



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 2
Flowsheet of the EBARA TwinRec process

- Volatile metal salts are concentrated into the secondary fly ash and can be used for zinc, lead, and copper recycling in the zinc industry.
- The amount of final residues for landfilling is reduced to very low values.

The energy efficiency of TwinRec is better than the thermal waste treatment processes that require oxygen and therefore internal consumption of electricity. Also, the ash-melting furnace is integrated into the water-steam cycle, making use of the highest temperature level for steam production.

The slag granulate is the largest fraction for recycling. For successful application in the construction industry, it must satisfy technical criteria and pass the respective environmental certification. Technically, the granulate qualifies for various applications, replacing cullet, gravel, or sand. It can be applied as loose bulk

material or as filler in combination with inorganic or organic binders. In Japan, the granulate is also used as a filler in asphalt.

Commercial Operational Experience

The first TwinRec commercial plant for MSW was built for Sakata Clean Union. The Sakata plant has a capacity of 2 x 98 t of MSW per day. Since the start-up of the first plant, several other TwinRec plants have been started, resulting in 15 plants in operation to date. Twelve of these plants treat MSW and are listed in Table 1.

Figure 3 shows a photograph of the largest plant at Kawaguchi that treats 420 t of MSW per day in three process lines generating 12 MW. In addition to vitrification of its own ashes, bottom and fly ash of another grate-type incinerator is also vitrified in the ash-melting furnace. Additionally, some of the

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 1 List of operational TwinRec plants

No.	Customer	Location	Capacity	Type	Inst. year	Electricity kw	
<i>Municipal waste</i>							
1	Joetsu Union	Niigata	15.7 t/24 h	TIFG	Mar.2000	45	Night soil sludge
2	Sakata Clean Union	Yamagata	196 t/24 h	TIFG	Mar.2002	1,990	
3	Kawaguchi City	Saitama	420 t/24 h	TIFG	Nov.2002	12,000	
4	Ube City	Yamaguchi	198 t/24 h	TIFG	Nov.2002	4,100	
5	Chuno Union	Gifu	168 t/24 h	TIFG	Mar.2003	1,980	
6	Minami-Shinshu Union	Nagano	93 t/24 h	TIFG	Mar.2003	800	
7	Nagareyama City	Chiba	207 t/24 h	TIFG	Feb.2004	3,000	
8	Chubu Clean Union	Shiga	180 t/24 h	TIFG	Mar.2007	3,000	
9	Dalsung	Korea	70 t/24 h	TIFG	Jun.2008	–	(HEEC license)
10	Eunpyeong	Korea	48 t/24 h	TIFG	Sep.2009	–	(HEEC license)
11	Hwasung	Korea	300 t/24 h	TIFG	Mar.2010	4,400	(HEEC license)
12	Kurahama Clean Union	Okinawa	309 t/24 h	TIFG	Mar.2010	6,000	
<i>Industrial waste</i>							
1	RER aomori renewable energy recycling Co., Ltd.	Aomori	450 t/24 h	TIFG	Nov-02	17,800	Shredder dust, sludge
2	Nikko Mikkaichi recycling Co. Ltd.	Toyama	63 t/24 h	TIFG	Jun-01	–	Shredder dust, waste plastic
3	Tokyo waterfront recycle power Co., Ltd.	Tokyo	550 t/24 h	TIFG	Aug-06	23,000	Industrial waste



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 3
Photograph of the Kawaguchi plant

secondary fly ash is recirculated and even the inert gasifier bottom ash, after metals separation, is ground and fed back to the ash-melting furnace. In this way, over 97% of the waste input is transformed into energy, metals, and recyclable glass granulate.

Figure 4 shows a photograph of the Tokyo Water-front Recycle Power plant located in Tokyo Bay area, and treating 22.9 t/h of industrial waste in two process lines. In this plant, industrial wastes (plastic wastes and crushed/separated residue of construction wastes) are received in shredded form. Ash is melted under high temperatures into slag that is granulated and used as construction material. This plant generates 23 MW of electricity by recovering the heat generated in this plant and in another facility, next to this plant, in which medical wastes are treated.

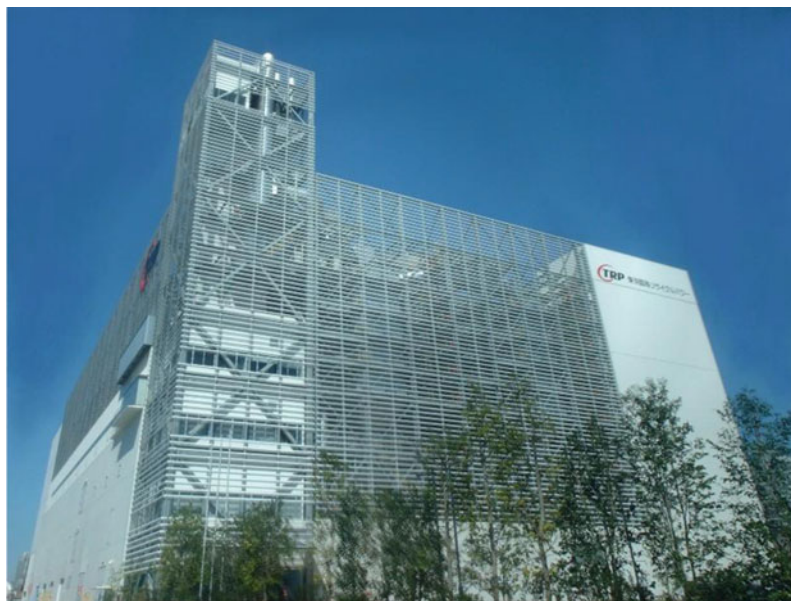
The JFE High-Temperature Gasifying and Direct Melting Process

Process Description

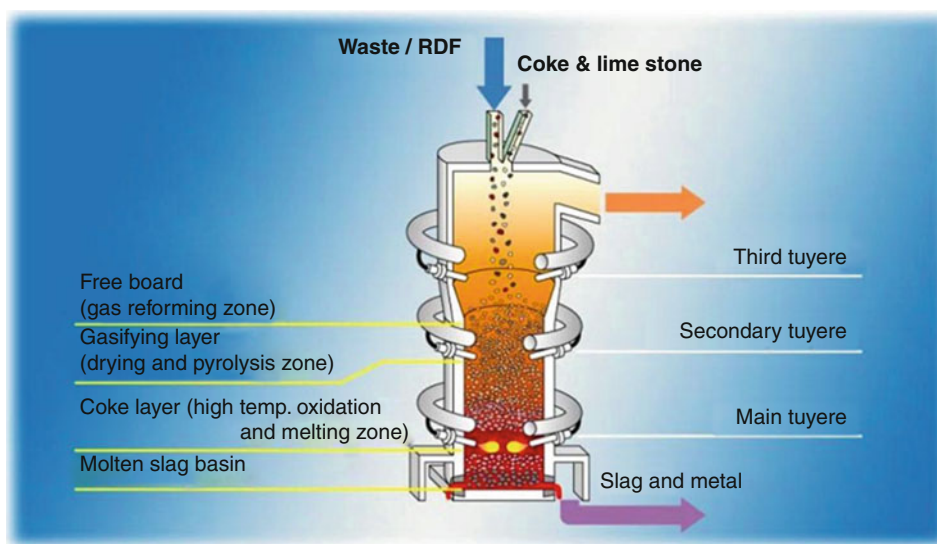
JFE is a new company resulting from the merger of Nippon Kokan (NKK) and Kawasaki Steel. The JFE High-Temperature Gasifying and Direct Melting

Process (JFE Process) resembles a small iron blast furnace where wastes are fed through the top of a vertical shaft (Fig. 5).

Air is introduced into the furnace through primary, secondary, and tertiary tuyeres located along the height of the shaft. The primary air, near the bottom of the shaft, is enriched to about 35% oxygen in order to generate the high temperatures required to transform the ash to molten slag and metal. In the gasifying zone, the gas produced in the lower part is partially combusted at approximately 600°C by an air sent through the secondary tuyeres while maintaining a fluidized state. By means of this heat, the wastes charged from the furnace top are preheated and thermally decomposed. Also, the fluidization ensures the downward flow of the bed within the shaft. In the gas reforming zone (“free-board”), a tertiary air flow is injected to maintain the freeboard outlet temperature at 850°C and decompose organic gases and tar in reducing atmosphere. Ample space in the free board stabilizes the gas flow and reduces the velocity resulting in lower dust carryover in the gas flow. The slag and metal overflow from the furnace are quenched in a water



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 4
Photograph of the Tokyo waterfront recycle power plant



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 5
JFE High temperature gasifying and direct melting process

tank to form small spherical particles of granulated slag and metal.

The process requires the addition of coke (less than 5% of wastes), which is also added at the top of the

shaft along with sufficient lime to form a fluid slag at the bottom of the furnace. The JFE Process produces slag and metal globules that are used beneficially, and fly ash which contains volatile metals and is landfilled.

Commercial Operation Experience

Up to 2010, JFE has delivered ten Direct Smelter plants in Japan, as shown in Table 2 [2]. All of them process as-received MSW except for the Fukuyama plant where RDF is combusted. The most recent plant serves the Chikushino/Ogori/Kiyama association in Fukuoka Prefecture (Kyushu) introduced. This plant is called “Clean-Hill Homan” and will be described in the following sections.

An Example of the Performance of the JFE Direct Melting Process

Figure 6 shows the process flowsheet and Fig. 7 is a photograph of the most recent JFE Direct Smelting plant at Fukuoka.

Table 3 shows the principal components of the Fukuoka plant.

Table 4 shows the mass balance of this plant in 2008. The total weight of MSW treated was 49,348 t and the slag, metal, and fly ash were 11.1%, 0.7%, and 2.3% of the solids feed, respectively. The use of the cyclone shown in Fig. 6 reduced the amount of fly ash significantly. All the slag recovered was utilized as a secondary

concrete material or as a sub-base in road construction. The metal and fly ash recovered were also recycled.

Table 5 shows the electric power balance including the power usage in the recycling center of this plant. The 49,000 t of MSW generated 22,000 MWh of electricity of which 9.100 MWh were sold to the grid.

Table 6 shows the exhaust gas emission data along with the national standards. All the emission data were well below the regulation values.

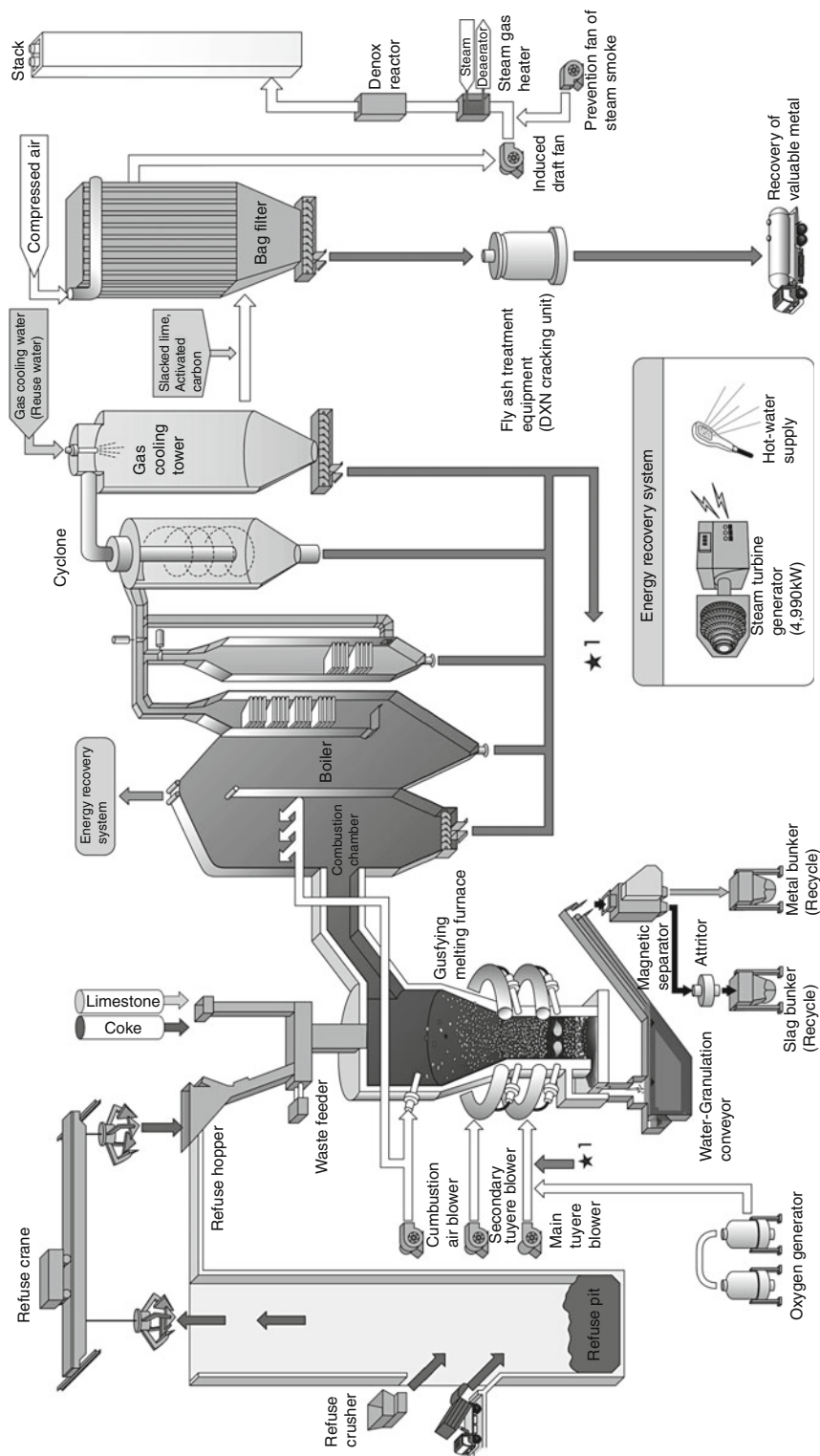
Table 7 compares the results of leachability and concentration tests for various metallic contaminants with the standard values of Japan. It can be seen that in all cases the test data were substantially below the standard values.

Slag Utilization

Japanese government has a policy of encouraging the vitrification of ash ((bottom ash to slag) as part of the hierarchy of waste management and for extending landfill life. Therefore, the production of slag has been increased remarkably during the last 10 years. Slag is standardized by JIS (Japan Industrial Standard) for usage as asphalt and concrete aggregate. As a result, slag utilization is progressing and a considerable

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 2 List of operational JFE process plants

	Municipality/owner	Capacity × Line	Input waste	Completion
1	Kagamihara City, GIFU	192 t/day (64 t/day × 3)	MSW (incl. bulky wastes)	2003.03
2	Amagi/Asakura/Mitsui Association, FUKUOKA	120 t/day (60 t/day × 2)	MSW (incl. bulky wastes)	2003.03
3	Hidaka-chubu Association, HOKKAIDO	38 t/day (19 t/day × 2)	MSW (incl. bulky wastes)	2003.02
4	Morioka/Shiwa Area Association, IWATE	160 t/day (80 t/day × 2)	MSW (incl. bulky wastes)	2003.03
5	Saiki Area Association, OITA	110 t/day (55 t/day × 2)	MSW (incl. bulky wastes)	2003.03
6	Fukuyama Recycle Power Corp., HIROSHIMA	314 t/day (314 t/day × 1)	RDF	2004.02
7	Ibaraki Environment Protection Foundation, IBARAKI	145 t/day (72.5 t/day × 2)	MSW and industrial waste (incl. bottom ash)	2006.03
8	Aki Area Association, KOCHI	80 t/day (40 t/day × 2)	MSW (incl. bulky wastes, landfill-wastes)	2006.03
9	Hamada Area Association, SHIMANE	98 t/day (49 t/day × 2)	MSW (incl. bulky wastes)	2006.11
10	Chikushino/Ogori/Motoyama Association, FUKUOKA	250 t/day (125 t/day × 2)	MSW (incl. bulky wastes, disaster wastes)	2008.03



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 6
Process flowchart of the Clean-Hill Homan plant



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 7
Photograph of the Clean-Hill Homan plant

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 3 Outline of the clean-hill homan plant

Capacity	250 t/day (125 t/day \times 2 Furnaces)
Furnace type	Shaft melting process (high temperature gasifying and direct melting furnace)
Energy recovery system	Boiler 22.0 t/h, 400°C \times 3.92 MPa, steam turbine generator (4 990 kW), hot-water supply system
Exhaust gas treatment system	Cyclone, bag filter, Denox reactor
Slag treatment system	Water-granulation conveyor, magnetic separator, attritor
Fly ash treatment system	Dioxins cracking unit

fraction of slag has acquired an economic value. Figure 8 shows the increase in number of ash-melting furnace plants with time. Figure 9 also shows that both the tonnage of slag produced and the slag used beneficially have increased with time.

The various uses to which the slag is put are shown graphically in Fig. 10.

The TOSHIBA Process for Liquefaction of Plastic Wastes

Social Background

The Plastic Waste Management Institute of Japan reported [3, 4] that the domestic plastic waste produced in 2006 had reached a total of 10 million tons, made up of about 5 million tons of household waste and another 5 million tons of industrial waste. Of this waste, 72% (7.21 million tons) was reutilized as materials, fuels, electricity, or heat, among others. However, 28% (2.84 million tons) was incinerated without energy recovery or landfilled. According to the Japan Containers and Packaging Recycling Association [5], the quantity of plastic containers and wrapping, within household plastic wastes was 550,000 t. Of this amount, 23% (130,000 t) was used in materials recycling operations and 46% (250,000 t) in chemical recycling operations, under the Container and Packaging Recycling Law; the remaining 31% was incinerated or landfilled. The breakdown of chemical recycling activities (250,000 t) in 2006 were coke ovens (61%), gasification (22%), blast furnaces (15%), and liquefaction (2%). The waste plastics liquefaction operations of the Sapporo Plastics Recycling Co., Ltd. (SPR) are classified as a chemical recycling technique in Japan. Two plastic

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 4 Mass balance of MSW disposal (total MSW disposal was 49,348 t)

Recovered material	Amount of emergence (t)	Ratio (wt%)
Slag	5,502	11.1
Metal	352	0.7
Fly ash	1,116	2.3

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 5 Electric power balance (Including recycling center)

Item	Electric power (MWh)
Generated	22,349
Purchased	989
Sold	9,070
Consumed	14,268

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 6 Exhaust gas emission data

Item	Regulation value	Analysis value
Dust (g/m ³ N)	<0.02	<0.005
SO _x (ppm)	<50	0.3–6.0
NO _x (ppm)	<50	6.0–32.0
HCl (ppm)	<50	<8.6
CO (ppm)	<30	2–7
Dioxins (ng-TEQ/m ³ N)	<0.05	0.00000009–0.0048

liquefaction facilities have been operating commercially in Japan: the Niigata Plastics Liquefaction Centre (6,000 t/year) and the Sapporo Waste Plastics Liquefaction Plant (14,800 t/year). The waste plastics liquefaction technique is different to other recycling techniques, and, after overcoming initial problems,

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 7 Result of slag measurement (example)

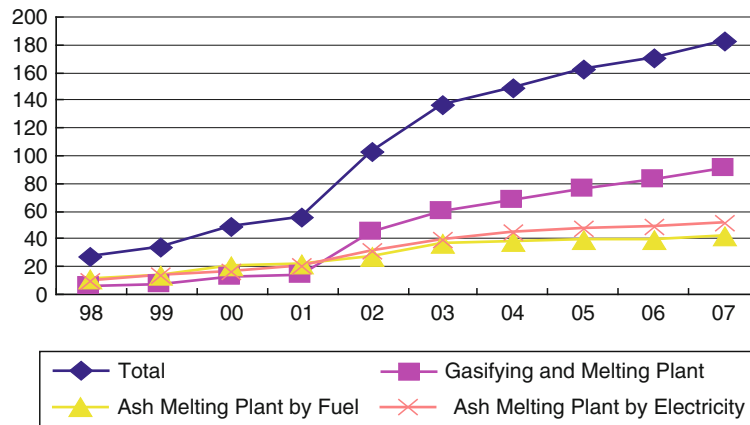
	Elution (standard value, mg/l)	Content (standard value, mg/kg)
Cd	<0.005 (<0.01)	<10 (<150)
Pb	<0.005 (<0.01)	<10 (<150)
Cr ⁶⁺	<0.04 (<0.05)	<10 (<250)
As	<0.005 (<0.01)	<10 (<150)
T-Hg	<0.0005 (<0.005)	<0.1 (<15)
Se	<0.005 (<0.01)	<10 (<150)
F	<0.08 (<0.8)	<150 (<4,000)
B	<0.1 (<1.0)	<150 (<4,000)

SPR process has maintained high levels of safety, stability, and productivity.

Process Description

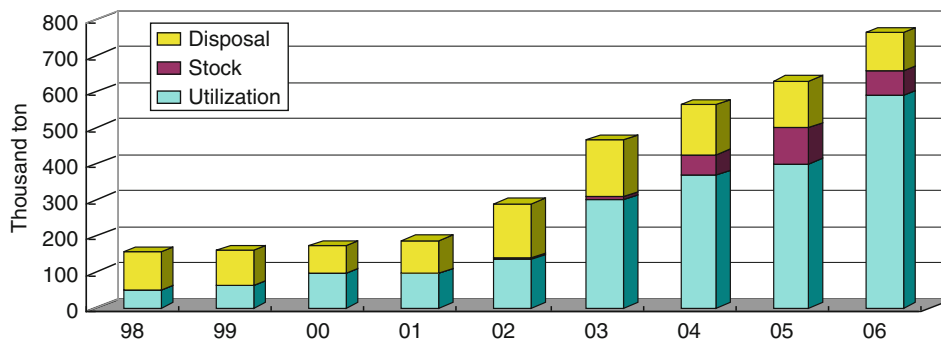
In 2000, SPR started operating a liquefaction process that includes the unique characteristic of dechlorination of plastic wastes that contain polyvinylchloride (PVC). The flowsheet of the SPR plastic waste liquefaction process is shown in Fig. 11. In the pretreatment stage, bales of compacted waste plastics are shredded and then foreign materials, such as pieces of metal and water are removed, and the remaining plastics are pelletized. The pellets are then fed into the dechlorination process where they are heated electrically to 300–330°C, melted, and the hydrochloric gas resulting from the thermal decomposition of PVC is incinerated at 1,200°C in dechlorinating furnace; scrubbing of this gas yields a solution containing less than 20% HCl which is sold. The molten polymer that is obtained in the dechlorination process is fed into the pyrolysis reactor where it is heated at 400–450°C for about 10 h and separates into a gaseous product that is conveyed to the distillation process and a residue that is fed to the solid fuel production process.

The gaseous product of the pyrolysis reactor is liquefied by spray quenching at 120°C, and the resulting pyrolysis oil is fed into the distillation tower where it is separated into three fractions: light oil at 120°C, “medium” oil at 200°C, and heavy oil at 280°C.



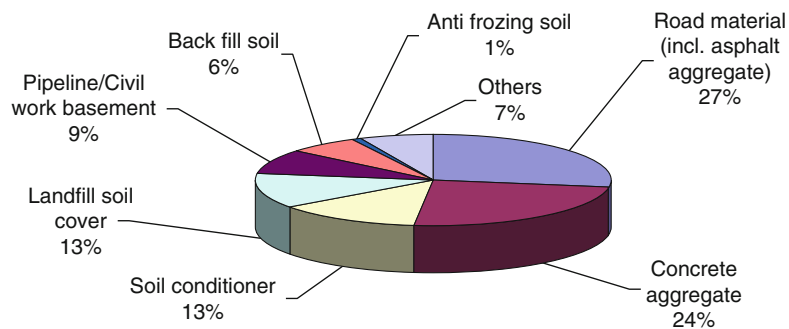
Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 8

Increase in number of ash melting furnace plants



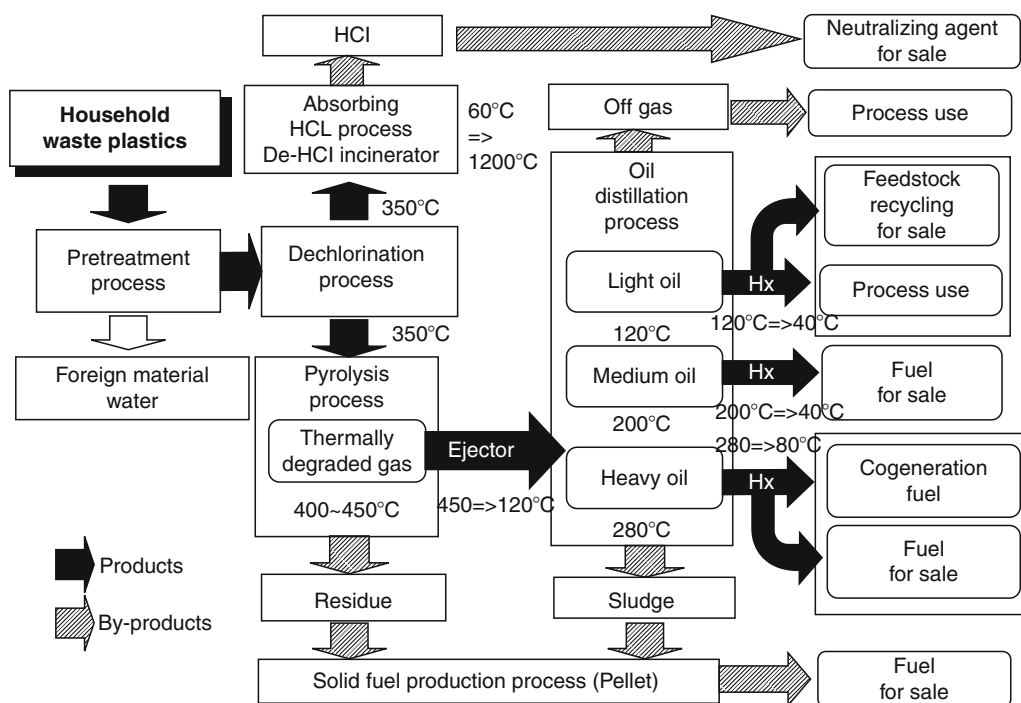
Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 9

Increase of slag production and beneficial use



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 10

Beneficial uses of slag



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 11

Flowsheet of the Sapporo plastics recycling (SPR) plastics liquefaction process. Hx Heat exchanger

The remaining volatile hydrocarbon gas is used as fuel in the plant operation. Some of the light oil product is used as a raw material for manufacturing plastic and the rest is used as fuel in the plant. The “medium” oil is sold to local companies and used as boiler fuel. Some of the heavy oil product is provided to local central heating and air-conditioning companies, paper manufacturing companies, and other companies and used as fuel, and the rest is used to power cogeneration diesel engines. The sludge residue derived from the filtering of the heavy oil is mixed with pyrolysis residues and used as a solid fuel. Almost all the plastic, except for the foreign material and water, is being reclaimed. As a result, in 2006, the recycling rate, excluding water contained in the feedstock bales, reached 96%.

Main Technical Challenges

In the first year of operation (2000), it was difficult to maintain normal processing due to corrosion and clogging due to the presence of polyethylene terephthalate (PET) in the plastic waste. The

operational problems were due to the formation of benzoic acid (C_6H_5COOH) and terephthalic acid $C_6H_4(COOH)_2$, during the thermal decomposition of PET (Table 8). The cause of the problems was investigated and it was found that the organic acids were mainly formed in the operating temperature range between 170°C and 250°C. This problem was solved by adding hydrated lime [$Ca(OH)_2$] to the plastic waste pellets and plant operation was stabilized.

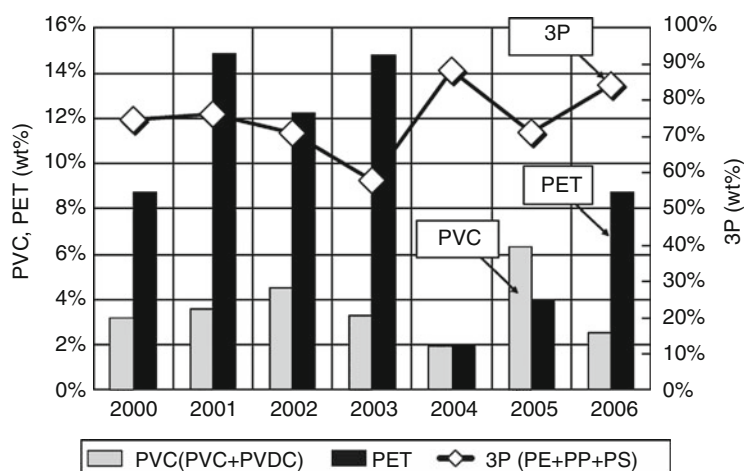
Composition of Raw Material and Properties of Reclaimed Products

The composition of typical municipal plastic waste is shown in Fig. 12. Polypropylene, polyethylene, and polystyrene (PP/PE/PS:3P), which are easily processed by liquefaction, make up 70–90% of the waste. However, PVC and PET constitute 2–7 t% and 2–15% of the waste, respectively. PVC causes corrosion and quality degradation of the recycled product, and PET can cause corrosion and clogging. Because the problems caused by corrosion and

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 8 Influence of organic acids on the recycling process

	Temperature (°C)		Phase	Concentration (wt%)		Trouble (•; Serious)	
	In	Out		Benzoic acid	Terephthalic acid	Clogging	Corrosion
Heavy oil Hx (upper)	250	170	Liquid	<27	<1	•	None
Heavy oil Hx (lower)	170	80	Liquid	–	–	Minor	None
Fire heater tube	200	300	Liquid			None	•
Distillation tower (upper middle column)	210		Liquid	<90	<2	None	•
Distillation tower (middle column)	220		Liquid	<24	<15	None	•
Distillation tower (bottom column)	270		Liquid	<7	<3	Minor	None

Hx Heat exchanger, wt% weight percent

**Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 12**

Composition of municipal waste plastics. *PET* polyethylene terephthalate, *PVC* polyvinylchloride, *PVDC* polyvinylidene chloride, *PE/PP/PS* polyethylene, propylene, polystyrene

clogging were overcome by the countermeasures mentioned above, the operation of the SPR process is presently stable and safe.

The properties of the reclaimed oil are shown in Table 9. The sulfur, nitrogen, and chlorine contents were below the limit values specified in Japanese Industrial Standards (JIS; technical specification Z 0025 for pyrolytic oil from waste plastics). The SPR reclaimed oil contains 0.003–0.08% sulfur (JIS level: 0.2%), 0.08–0.14% nitrogen (JIS level: 0.2%), and 50–70 ppm chlorine (JIS level: 100 ppm).

The properties of the solid fuel produced from waste plastics are shown in Table 10. Solid fuel pellets are produced from thermally degraded residue and heavy oil sludge. Inorganic chloride levels of 1–4% are found in the solid fuel because CaCl_2 is formed by reactions between hydrated lime $[\text{Ca}(\text{OH})_2]$, as noted above is added to the process to mitigate the problems caused by PET and chlorine. Therefore, this solid fuel is used in a blend with other solid fuels (e.g., wood or coal) at low levels of a few percent (usually below 5%) to minimize environmental problems.

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 9 Properties of the oils recovered in the SPR process

Property		Light oil	Medium oil	Heavy oil	JIS TS Z0025 (Japanese technical standard)
Density	g/cm ³ (15°C)	0.814	0.824	0.856	
Flash point	°C	<21	78	114	
Pour point	°C	<−50	−35.0	47.5	
Reaction	pH	Neutral	Neutral	Neutral	
Ash	wt%	<0.001	<0.01	<0.01	≤0.05
Sulfur	wt%	0.002	0.03	0.08	≤0.2
Nitrogen	wt%	0.08	0.14	0.1	≤0.2
Chlorine	wtppm	50	70	60	≤100
Gross heating value	kJ/kg	42,070	45,040	45,360	

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 10 Properties of the solid fuel produced in the SPR process

Property	Solid fuel (eco pellet)	Degradation residue	Sludge
Lower heating value (kJ/kg)	15,160	17,570	31,650
Carbon (wt%)	41.8	45.6	67.5
Sulfur (wt%)	0.09	0.06	0.04
Nitrogen (wt%)	0.43	0.4	0.39
Hydrogen (wt%)	5.7	2.6	6.3
Calcium (mg/kg)	88,000	125,000	46,700
Silicon (mg/kg)	12,000	52,300	19,800
Aluminum (mg/kg)	13,000	12,700	940
Total chlorine (wt%)	2.82	4.96	1.42
Inorganic chlorine (wt%)	2.82	4.79	1.42
Bulk specific gravity (kg/l)	0.72	0.389	0.868

The measured results for gas emissions from the SPR off-gas-fired furnace are shown in Table 11. Some of the light oil is used as in-plant furnace fuel and the heavy oil for powering the cogeneration engines.

According to periodic analysis of the gaseous emissions of these processes, nitrogen oxides (NO_x), sulfur oxides (SO_x), dust, dioxins, and HCl levels are below the required standards.

Development of Recycled Products and By-Products

In the SPR process, sludge is separated from the heavy oil by a centrifugal filtering method that was installed after initial operation; this resulted in much better quality of the heavy oil product of this process and, since then, heavy oil has been sold to other companies for use as a fuel. The light oil has been sold to a petrochemical company and is used as raw material for the production of naphtha since 2004. It is also used in the production of plastics. The Japanese recycling law considers only hydrocarbon oil as a recycled product; thermal degradation residue, off-gas (flammable gas), and hydrochloric acid are not considered as recycled products.

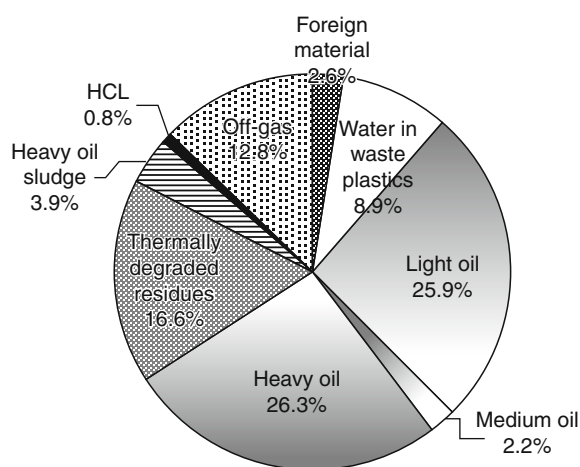
The off-gas of the SPR process has been reused as fuel within the plant since 2000. Initially, most of the thermal decomposition residue and oil sludge were discarded as industrial waste, but since 2004 they are supplied to other companies and are used as solid fuel. Hydrochloric acid, after neutralization, was initially discharged into the sewage system, but since 2004, it is also used by other companies as a neutralizer.

The actual recycling rate of the SPR plastics liquefaction plant in 2006 is shown in Fig. 13.

Gasification and Liquefaction Alternatives to Incineration in Japan. Table 11 Properties of gas emitted from the off-gas-fired furnace of the SPR waste plastics liquefaction plant

Periodic survey (twice a year)	Result of measurement	Emission standard	Date
Dust (particulates)	$<0.02 \text{ g/Nm}^3$	0.15 g/Nm^3	2007/11/21
Sulfur oxide (SOx)	$<0.05 \text{ Nm}^3/\text{h}$	$3.12 \text{ Nm}^3/\text{h}$	2007/11/21
Nitrogen oxide (NOx)	95 vol ppm	150 vol ppm	2007/11/21
Optional survey			
Hydrogen chloride (HCl)	2.3 mg/Nm^3	80 mg/Nm^3	2004/1/13
Dioxin	$0.000018 \text{ ng-TEQ/Nm}^3$	$0.0006 \text{ ng-TEQ/Nm}^3$	2004/1/20

TEQ Toxicity equivalency quantity, Nm^3 Gas volume at 1 atm and 0°C



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 13

Actual recycling rate of the SPR plant in 2006

The recovered hydrocarbon oil amounted to 54.4% of the weight of the initial waste plastic; the gaseous fuel, solid fuel, and the hydrochloric acid products amounted to 35% of the weight of the initial waste plastic. Since SPR recycles almost all of the input waste plastics, except for the water and the foreign materials, a high recycling rate of 96%, excluding the water content, has been achieved (Fig. 13). Also, most of the recovered materials are reused in the local Hokkaido district; as a result, the resource recovery rate for local communities in Hokkaido has reached 93%.

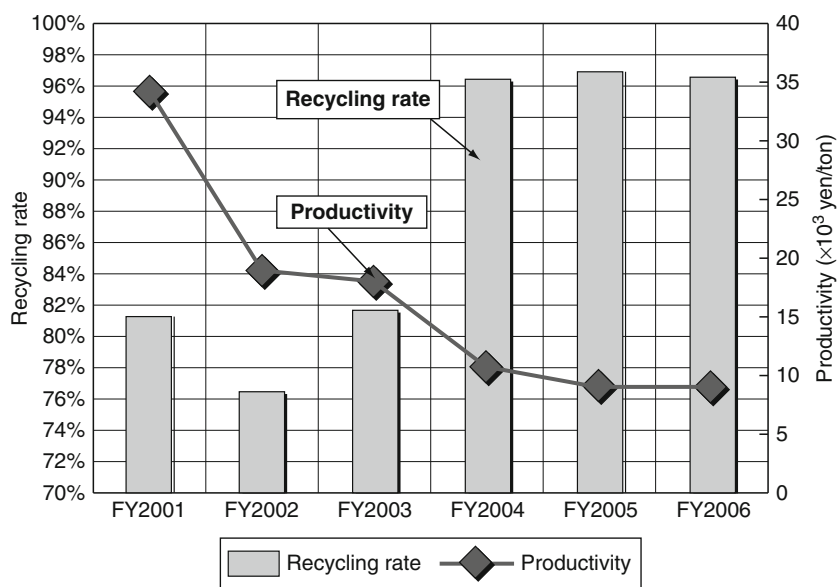
Productivity Improvement

The change in recycling rate and productivity with time of the SPR process are shown in Fig. 14. Productivity is expressed as the total operating cost of electric power, commercial fuel oil, and other supplies, etc.; these costs have decreased by a factor of three since the plant started operations in 2000. The measures that have contributed to productivity improvement are as follows:

- Reduction of hydrocarbon oil consumption during processing
- Reduction of amount of water used by producing and selling hydrochloric acid for use as neutralizer
- Reduction of industrial waste volume by selling solid fuel to others
- Introduction of an energy-saving burning system

Summary

Through its technological improvements and operational know-how, SPR has been able to process municipal waste plastics of almost all quality grades, even those containing PVC and PET. SPR can also process the sorted waste plastics from material recyclers or mechanical recyclers, thus allowing for a more efficient recycling of plastic waste that combines mechanical and chemical recycling. SPR has achieved an extremely high recycling rate (93% in 2006) from mixed plastic wastes and developed a system that allows for the use of the light oil product as a petrochemical raw material.



Gasification and Liquefaction Alternatives to Incineration in Japan. Figure 14
Annual trends of actual recycling rate and productivity

Future Directions

In direct comparison with the currently more common stoker grate incineration of MSW, the EBARA TwinRec and the JFE Direct Smelting processes offer a number of advantages: high recovery of metals and inert materials directly from the bottom ash and vitrification of fine ash particles into an inert construction material [1, 3].

These processes are based on gasification and require a lower amount of excess air, resulting in a compact air pollution control system. Also, as shown by the feedstock of the reference plants noted above, these processes are more flexible with regard to feedstock.

The processes described in this entry have demonstrated, through the range of capacity of commercial plants and the use of multiple feedstocks, that gasification of solid wastes is mature, reliable, efficient, and a good solution for current and future waste management applications.

Another approach to enhance the recycling of MSW is to segregate plastic materials and liquefy them, as demonstrated commercially by the SPR process. In the past it was difficult to recycle municipal waste plastics that contained PVC and PET. Thus, material

recycling methods have traditionally sorted out only PP/PE/PS from municipal waste plastics and nearly one half were disposed as residue. The SPR commercial liquefaction plant has shown that it is possible to process mixed plastics containing PVC and PET. This plant has attained a very high recycling rate achieved a high recycling rate of 93% of the solids in the feedstock to the process.

Bibliography

1. Selinger A, Steiner Ch, Shin K (2003) TwinRec gasification and ash melting technology – now also established for municipal waste. Presented at 4th international symposium on waste treatment technologies, Sheffield, 29 June–2 July 2003
2. JFE High Temperature Gasifying & Direct Melting System (2010) Operational result of clean-hill human municipal solid waste treatment center, JFE Technical Report, No. 25, pp 70–71, -(in Japanese)
3. Fukushima M, Shioya M, Wakai K, Ibe H (2009) Toward maximizing the recycling rate in a sapporo waste plastics liquefaction plant. *J Mater Cycles Waste Manag* 11:11–18
4. Plastic Waste Management Institute <http://www.pwmi.or.jp/flow/flame.htm> (inJapanese)
5. The Japan Containers and Packaging Recycling Association <http://www.jcpra.or.jp/eng/statistics.html>

Genetic Engineering of Crops for Insect Resistance

JOHN A. GATEHOUSE

School of Biological and Biomedical Sciences, Durham University, Durham, UK

Article Outline

Glossary

Definition of the Subject

Introduction

Insecticidal Proteins from *Bacillus thuringiensis*

How Do *Bt* Toxins Work?

Expression of Genes Encoding *Bt* Insecticidal Proteins in Transgenic Plants

Taking Transgenic Plants Expressing *Bt* Toxins into the Field

Developments to “First Generation” Crops Expressing *Bt* Toxins

Exploitation of Endogenous Plant Defensive Mechanisms Against Insect Herbivores

Some Novel Approaches

Insect-Resistant Genetically Engineered Crops and Sustainability

Future Directions

Bibliography

Glossary

Coding sequence The part of a gene which determines the sequence of the protein product

Domain A region of a protein which forms a distinct 3-D structure, and will often form this structure even when separated from the rest of the protein

Genetic engineering Introduction of a specific DNA sequence into an organism by artificial means

Insect orders Lepidoptera=butterflies and moths; diptera=flies; coleoptera=beetles; hemiptera/homoptera=sucking insects such as aphids

Mutagenesis Alteration to a DNA sequence, often resulting in alteration to the sequence of a protein which the DNA specifies

Oligomerization Formation of polymers containing a relatively low number of repeating units

Proteolysis Introduction of breaks in the chain of amino acids making up a protein by a proteinase

Transgenic Organism into which a gene has been introduced by genetic engineering technology

Definition of the Subject

Genetic engineering of crops for insect resistance is the introduction of specific DNA sequences into crop plants to enhance their resistance to insect pests. The DNA sequences used usually encode proteins with insecticidal activity, so that in plants which contain introduced DNA, an insecticidal protein is present. However, other strategies to improve plant defenses against insects have been explored. Genetically engineered crops that are protected against major insect pests by production of insecticidal proteins from a soil bacterium, *Bacillus thuringiensis*, have become widely used in global agriculture since their introduction in 1996.

Introduction

Twenty years have elapsed since the first publications describing transgenic plants, which showed enhanced resistance to insect herbivores, as a result of the expression of a foreign gene encoding *Bacillus thuringiensis* (*Bt*) toxin [1–3]. In the intervening years, crops expressing these toxins have become widely used in global agriculture, and have led to reductions in pesticide usage and lower production costs [4]. At the same time, the predictions made by lobby groups supporting “organic” crop production, that irreversible environmental damage would be caused by genetically engineered (GE) crops resistant to insect pests, have not been realized [5]. Despite all the controversy that GE crops have caused in many countries, it is difficult to dispute that the use of this technology to combat insect pests has had a positive impact on global agriculture.

This entry has two aims: first, to provide a summary of how and why *Bt* toxins have become the insect resistance genes of choice for commercial GE crop applications, and to anticipate some further developments of this technology; second, to consider some of the other approaches to engineering insect resistance in plants, and to assess their potential for future development in the development of sustainable agriculture.

Insecticidal Proteins from *Bacillus thuringiensis*

The presence of insecticidal toxins in the soil bacterium *Bacillus thuringiensis* (*Bt*) has enabled both the bacteria themselves, and genes derived from them, to be exploited as plant protectants. The toxicity is almost invariably based on proteins produced during sporulation of the bacteria, which form crystalline deposits associated with the spores. The insecticidal *Bt* proteins are encoded by genes present on plasmids, and the presence of these plasmids is the main feature which distinguishes *Bt* from other spore-forming bacilli [6]. Preparations of *Bt* spores have been used since the 1920s as a conventional, spray insecticide (and, as a “natural” product, are approved for use in organic agriculture), but their efficacy in the field is limited by inactivation and low persistence.

The ecological niche occupied by *Bt* appears to be simple to define. The life cycle starts with a spore and associated crystalline protein body which may be present in the soil. On being eaten by an insect, the protein deposit associated with the spore is dissolved and digested, converting the crystalline protoxin to an active toxin. The insect is then killed, and the carcass provides nutrients for the growing bacteria, which multiply rapidly. When the insect carcass is exhausted, the bacteria sporulate; the spores are dispersed, and the cycle recommences. However, this cycle is clearly too simplistic, as the target insects for *Bt* toxins are only rarely soil dwellers, and the dose of spores required to kill an insect larva is too large for dispersed spores to have much effect. Although *Bt* is widely distributed, levels of the bacterium in soils are generally too low to have any effect on insects, and spraying plants with spores does not result in persistent protection as a result of the establishment of a high bacterial population. The species has been described as an opportunistic pathogen, which has evolved the sporulation mechanism as a “backup” system to ensure its survival under unfavorable conditions [7]. *Bt* is naturally present in the phylloplane, as well as in soil, and has been detected on cabbage foliage [8], and in vegetative form on clover [9] at low levels, without any insecticidal effect. However, the insecticidal characteristic must be of benefit to the bacterium, since most of the insecticidal proteins are encoded by plasmids, and the plasmids are maintained in the *Bt* population as a whole, despite the obvious

metabolic costs of producing large quantities of spore-associated proteins. Not only are toxin-encoding plasmids maintained, but there is also a huge reservoir of diversity in the toxins themselves, and much effort has been put into screening bacterial isolates for strains of *Bt* with novel pesticidal activities [10].

Bt toxins are now classified on the basis of amino acid sequence similarity (an earlier classification system based on pesticidal activity has been superseded), in a systematic hierarchical system [11]. For the purposes of this contribution, only the major distinctions need be considered. There are four types of insecticidal proteins produced by *Bt*:

1. Proteins associated with *Bt* spores, usually as crystalline deposits; three domain structure; single toxins; designated by the symbol Cry
2. Proteins associated with *Bt* spores, usually as crystalline deposits; binary toxins and other similar proteins, including truncated versions of three-domain toxins; also designated by the symbol Cry
3. Proteins associated with *Bt* spores, usually as crystalline deposits; single domain structure; cytolytic; single toxins; designated by the symbol Cyt
4. Proteins expressed vegetatively by *Bt*; single chain and binary toxins; designated by the symbol Vip

Each type of toxin is subdivided (on the basis of sequence similarity) into families (number; same number $\geq 45\%$ sequence identity) and then further subdivided using capital letters (same letter $\geq 78\%$ sequence identity), small letters (same letter $\geq 95\%$ sequence identity) and numbers successively. The resulting system yields designations for specific toxins such as Cry1Aa. A single *Bt* strain can produce spores which contain only a single toxin, or a complex mixture, such as the *Bt* subspecies *israelensis*, whose spores contain Cry4Aa, Cry4Ba, Cry10Aa, Cry11Aa, Cyt1Aa, and Cyt2Ba toxins [12].

All four types of proteins have been proposed for use as crop protection agents, although Cyt toxins have not as yet been used in commercial insect-resistant transgenic plants, and three-domain Cry toxins are by far the most commonly used type. Cry and Cyt toxins belong to the class of proteins referred to as bacterial pore-forming toxins, and show structural similarity to the α -helical and β -barrel groups of toxins, respectively

(where α -helical and β -barrel refer to the structures of the membrane-spanning parts of the toxin; reviewed by Parker and Feil [13]). These pore-forming toxins show common features of activity; they are produced as water-soluble proteins, and interact with specific receptors on cell surfaces, often after proteolytic activation by host proteinases. Binding to cell surfaces triggers a conformational change leading to oligomerization, which allows insertion into the cell membrane through promotion of a fluid, partially denatured structure. Insertion of the toxin into the membrane can either cause cell death directly, or result in effects on intracellular metabolism which lead to cell death.

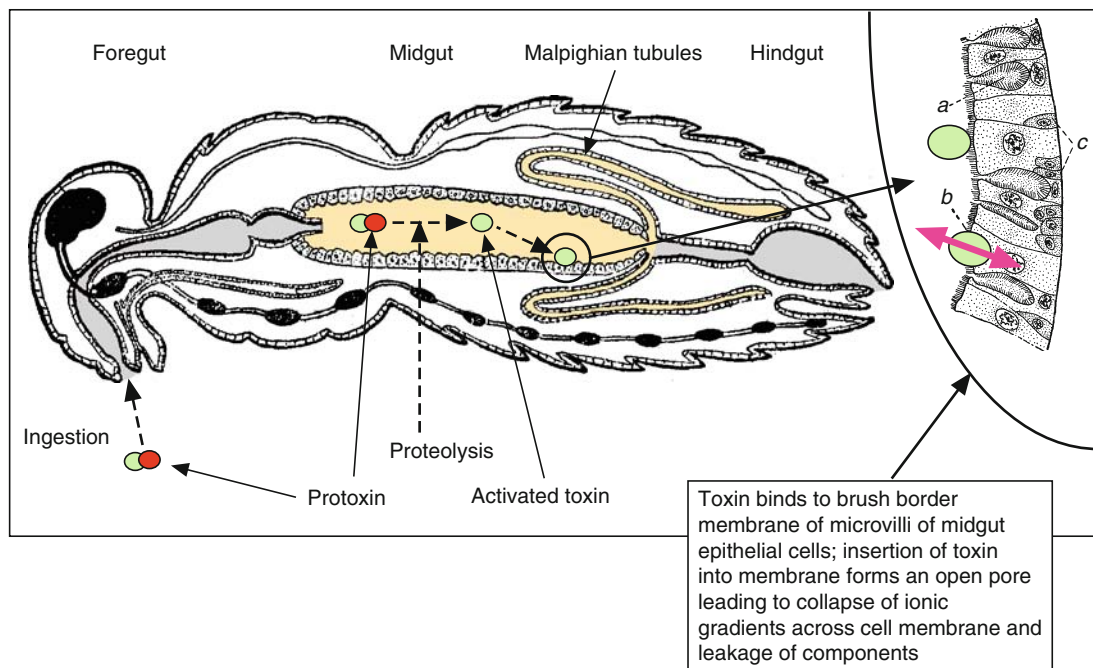
How Do *Bt* Toxins Work?

Three-Domain Cry Toxins

The mechanism of action of the “conventional” three-domain Cry toxins is now well understood, and can be divided into four stages:

1. Solubilization of the protoxin, and proteolytic activation by proteinases in the insect gut to produce active toxin
2. Interaction of the toxin with one or more receptors on cell surfaces in the insect gut epithelium
3. Oligomerization of the toxin
4. Insertion of the oligomerized toxin into cell membranes, leading to the formation of open pores, and cell death (see Fig. 1)

Following the pioneering work of Ellar's group [14] tertiary structures of six different three-domain Cry toxins are known – Cry1Aa [15], Cry2Aa [16], Cry3Aa [14], Cry3Bb [17], Cry4Aa [18], and Cry4Ba [19]; whereas most structures are for the active form, the structure of Cry2Aa includes the N-terminal pro-region. These toxins all show a high degree of structural similarity, and thus the formulation of a general model for their mode of action is justified. The three domains present in the active forms of these proteins are designated I, II, and III, and are normally contained in



Genetic Engineering of Crops for Insect Resistance. Figure 1

Action of *Bt* toxins on the insect gut epithelium. Death of insect results from disintegration of gut epithelium (due to cell death) and proliferation of gut microflora

a single polypeptide of approximately 600 amino acid residues (in some cases proteolytic cleavages are present within the active three-domain structure as a result of protoxin activation, resulting in multiple polypeptides making up the toxin, but the overall three-domain structure is conserved.). While conservation of structure and sequence is observed in the active forms of three-domain toxins, many toxins are synthesized with C-terminal extensions, which are variable in sequence between *Bt* strains, and in length between Cry families. The presence of C-terminal extensions leads to a large degree of heterogeneity in the size of the protoxins present in bacterial spores, with sizes ranging from approximately 600 amino acids (similar to the active toxin) to approximately 1,200 amino acids. These C-terminal extensions are not required for toxin function, and are removed during toxin activation, although their removal is not sufficient for toxicity to be shown. They are thought to play a role in the formation of crystalline inclusions in the bacterium during the spore-forming process.

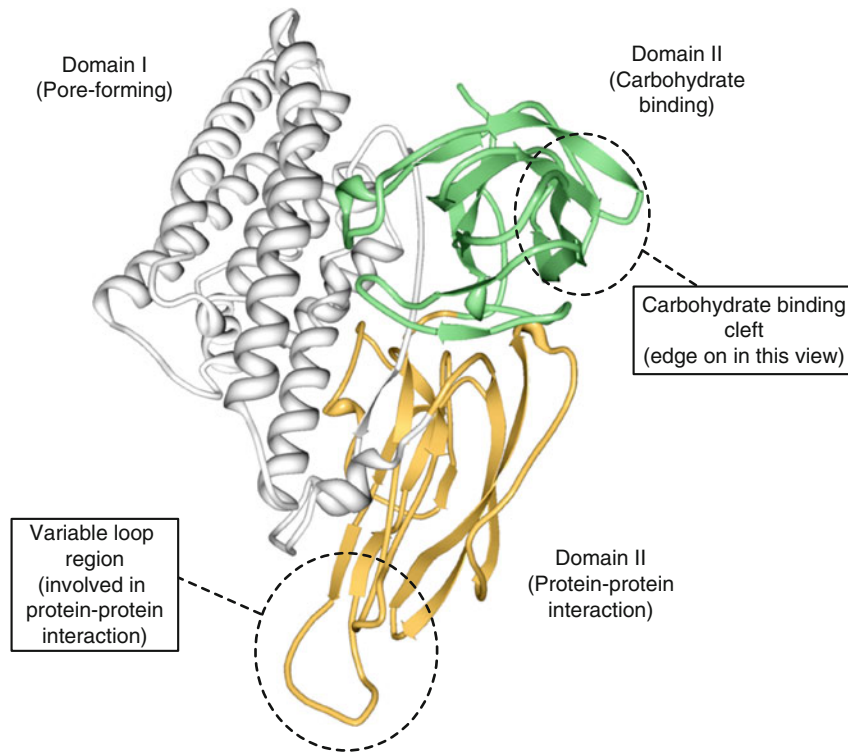
The three domains of the active toxin are clearly distinguished in their structures.

1. Domain I, approx. 260 aa, contains seven α -helices, of which six are amphipathic and one hydrophobic. This structure is typical of pore-forming toxins, with the hydrophobic and amphipathic helices being responsible for membrane insertion and pore formation. The hydrophilic sides of the amphipathic helices form the surface lining the pore, so that polar species such as ions are able to cross the membrane.
2. Domain II, approx. 170 aa, forms a “ β -prism” structure, with three β -sheets, and exposed loops on its surface.
3. Domain III, approx. 160 aa, has a compact structure with two anti-parallel β -sheets in a “jellyroll” formation, and is structurally similar to carbohydrate-binding domains such as the cellulose-binding domain in cellulases [20]. A general model for three-domain toxins is shown in Fig. 2.

The Proteolytic Activation Process Ingestion of the Cry protoxins by the insect leads to solubilization of the proteins, and exposure to digestive proteinases in the insect gut. Although removal of the C-terminal

protoxin region occurs at this stage, the essential step in protoxin activation is the proteolytic cleavage and removal of an N-terminal peptide, which varies from approx. 25–60 amino acids in different Cry proteins. A non-activatable Cry1Ac mutant toxin could not form pores in insect membrane vesicles derived from gut epithelial cells [21], and it is thought that the N-terminal peptide “masks” a region of the toxin involved with interaction with receptors [16]. The activated toxin is fairly resistant to further proteolytic cleavage, which enables it to survive long enough in the gut to reach its site of action, the gut epithelial surface (Fig. 1).

This summary overlooks a number of factors which contribute to toxicity. First, the location of the proteolysis may be important, since many insects, such as diptera (flies), carry out digestion in the foregut, which is chitin-lined and does not contain epithelial surfaces, or even outside the insect altogether, by secreted saliva or regurgitated gut contents. Under these circumstances, the toxin will need to be more resistant to proteolysis, or more effective, since the time between activation and reaching the site of action will be longer. Secondly, gut conditions vary significantly between insects from different orders, or even within orders; in general, larvae of lepidoptera (moths and butterflies) have a highly alkaline midgut environment (pH 10–11 in many major crop pests), whereas larvae of coleoptera (beetles) have an acidic gut environment (pH approx. 5 for many species). These differences in conditions will affect both the activation and survival of the protein, although they may be less relevant to steps taking place at the gut surface, where there is a separation from the gut lumen by the peritrophic membrane (a macroscopic porous chitin-based structure) and by lipids sloughed off from the gut surface. Finally, the nature of the digestive enzymes present in the insect gut differs considerably between different orders; whereas most insects use serine proteinases with an alkaline pH optimum as their major endoproteinases, many coleopteran larvae use cathepsin-type cysteine proteinases with an acidic pH optimum (similar to lysosomal proteinases). On the other hand, protoxin activation does not appear to be very sequence specific. Many lepidopteran-specific Cry proteins can be activated *in vitro* by mild treatment of the protoxin with bovine trypsin, yielding products that appear to be similar to those



Genetic Engineering of Crops for Insect Resistance. Figure 2

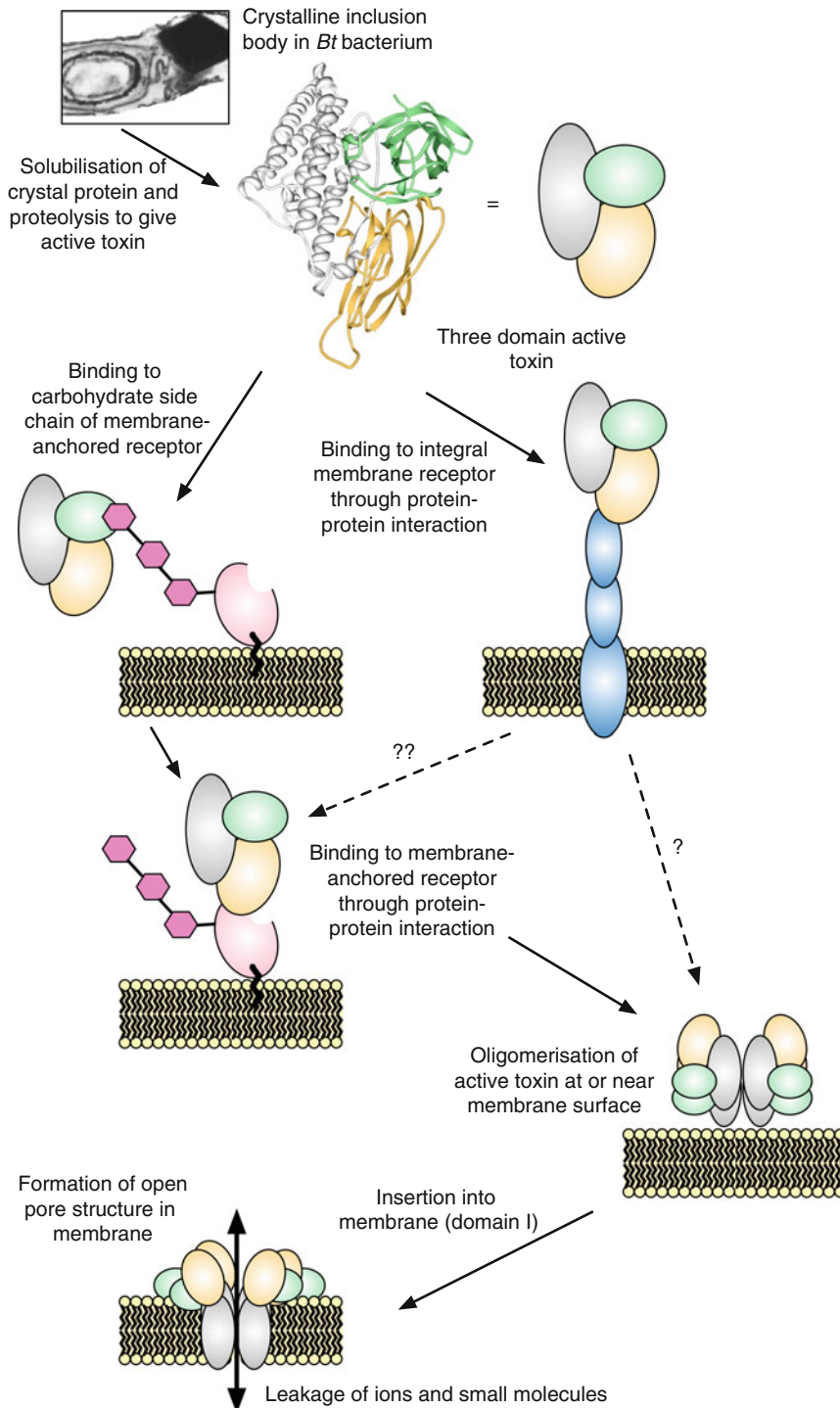
Model structure for three-domain *Bt* toxins. Ribbon diagram showing backbone structure of *Bt* toxin Cry1Aa (PDB 1ciy; [15]); structure of active toxin shown. The three domains are color coded: domain I, silver; domain II, orange, domain III, green. Features as shown on diagram

formed *in vivo*. This suggests that it is the three-dimensional structure of the protoxin that determines where proteolysis takes place, unless forcing conditions are used.

Interactions with Receptors Proteins to which Cry proteins bind in the insect gut are termed “receptors,” although the specificity of interaction is determined by the Cry protein itself, and the ligands to which it binds do not show the properties of receptors as normally understood. Binding takes place on the microvillar membranes of the cells forming the midgut epithelium, and involves interactions with relatively abundant proteins, either attached to the cell membrane by glycosylphosphatidyl-inositol (GPI)-anchors, or integral to the membrane with large extracellular domains. The overall process is summarized in Fig. 3.

Methods for identifying receptors to which Cry proteins bind have largely been based on immunoblotting of proteins prepared from brush border membrane vesicles (BBMV). This method is not a good mimic of conditions *in vivo*, and may result in interactions with lower affinity, or which are dependent on protein conformations maintained by membranes, not being observed. Nevertheless, the major binding partners for Cry proteins which have been identified show binding when assayed as purified proteins, and as components of BBMVs, with binding constants in the range 1–100 nM.

The initial identification of membrane-anchored aminopeptidase N [23] and an integral membrane cadherin-like protein designated *Bt*-R1 [24] as Cry1A toxin receptors in lepidopteran insects has been supplemented more recently by identification of a 270 kDa glycoprotein [25] and alkaline phosphatase



Genetic Engineering of Crops for Insect Resistance. Figure 3

Mechanism of action of three-domain *Bt* toxins. The scheme shown is adapted from the “two-receptor” model [22]

(membrane anchored; [26]) as additional potential receptors. Alkaline phosphatase appears to be the major receptor in mosquitoes [27]. A recent proteomic analysis has identified further potential receptors, such as V-ATP synthase subunit 1 [28]. However, this analysis also showed binding to actin, which could not be present at the cell surface, showing that results from blotting experiments need to be interpreted critically.

Functional roles as “receptors” for aminopeptidase N and cadherin *Bt-R1* in Cry protein toxicity are supported by numerous studies. Strains of lepidopteran insects resistant to Cry1 toxins have been identified which show mutations in the gene encoding *Bt-R1*, leading to the production of a truncated cadherin lacking the extracellular domains [29, 30]. The correlation with loss of function of cadherin with loss of susceptibility to Cry toxins suggests that binding to the extracellular domains of cadherin is a necessary step for toxicity. Binding of Cry1A toxin to the cadherin extracellular domains has been demonstrated in vitro, and the binding regions have been identified in some detail [31]. Both gain-of-function and loss-of-function assays have been used to provide further evidence for involvement of cadherin in toxicity; when transiently expressed in mammalian cells that were not normally susceptible to Cry toxin, *Bt-R1* genes from silkworm conferred sensitivity to Cry1A toxins [32]; whereas suppression of cadherin expression by RNA interference in tobacco hornworm (*Manduca sexta*) decreased sensitivity to Cry1Ab toxin [33]. In the case of aminopeptidase N, similar correlations between resistance to *Bt* toxin and lack of expression of specific isoforms of the protein have been observed [34], but more direct evidence has come from downregulation of aminopeptidase by RNA interference using double-stranded RNA. This technique has been carried out in lepidopteran larvae, giving decreased sensitivity to Cry1C toxin [35], and in lepidopteran cell cultures, giving decreased sensitivity to Cry1Ac [36]. A gain of function experiment in which transgenic fruit flies (*Drosophila*) expressing lepidopteran aminopeptidase N became sensitive to the lepidopteran-specific toxin Cry1Ac [37] showed elegantly and convincingly that this receptor plays a key role in toxicity. Binding to aminopeptidase N involves interaction of Cry toxins with the carbohydrate side-chains of the protein [38, 39], with specificity toward GalNAc residues being shown

(this sugar can inhibit binding; [40]). Binding to carbohydrate facilitates subsequent protein–protein interactions, which are thought to be necessary for toxicity [41]. Functional evidence for alkaline phosphatase acting as a Cry toxin receptor has again been provided by correlative observations, in that insect lines resistant to Cry1Ac toxins have lower alkaline phosphatase levels than susceptible lines [26]. Interactions with protein-bound carbohydrate also seem to be involved in the binding of Cry toxins to alkaline phosphatase.

The roles of the different domains of Cry proteins in the interaction with receptors are clearly distinguished. Despite the presence of the N-terminal propeptide which must be removed for activity, domain I plays little or no role in the interaction with receptors, whereas domain II is responsible for most protein–protein interactions (see Fig. 2), and domain III is responsible for binding to carbohydrates. This division of roles is consistent with the observation that a single toxin can interact with more than one type of “receptor”; for example, Cry1Ac interacts with both *Bt-R1* and aminopeptidase N [22]. The protein–protein interactions mediated by domain II have been localized to variable loop regions on the surface of the domain, whereas the carbohydrate-binding region of domain III is a typical binding site cleft, which is spatially well-separated from the domain II loops.

Oligomerization Oligomerization is a common feature in bacterial pore-forming toxins, and Cry proteins appear to conform to the model, with the formation of oligomeric structures (probably tetramers) observed for toxins from the Cry1 and Cry3 families. Mutants of the Cry1Ab protein that have impaired oligomerization ability, but bind to the receptor, show much reduced toxicity or no toxicity toward lepidopteran larvae [42]. Similarly, monomeric Cry proteins have much lower intrinsic pore-forming abilities on synthetic membranes than oligomerized preparations [43]. Oligomerization is promoted by binding to a receptor; in the case of Cry1Ab protein binding to the cadherin *Bt-R1* receptor, this process involves an additional proteolytic cleavage at the N-terminal end of the protein, in domain I [44]. The proteolytic cleavage, carried out by host enzymes, may aid the oligomerization process. The importance of oligomerization in promoting toxicity has been shown by two

complementary studies. First, a peptide corresponding to the region of cadherin to which Cry1A binds has been shown to act as a synergist, increasing the toxicity of Cry1A toward lepidopteran larvae [45], presumably as a result of the binding between the peptide and Cry1A promoting oligomerization of the toxin prior to interaction with the gut epithelium. Secondly, mutants of Cry1Ab toxin have been produced which contain deletions corresponding to the proteolysis in helix 1 of domain I which occurs on binding to cadherin. These mutated toxins form oligomers in the absence of cadherin binding, and are effective against insects that have cadherin expression suppressed, or which have a cadherin mutation which leads to resistance to unmodified toxin [33]. These results have led to a current view that cadherin is the primary receptor for Cry toxins, since it is necessary to promote oligomerization, with other molecules taking the role of “secondary receptors” [33].

Insertion into the Cell Membrane The oligomeric Cry protein must partially unfold in order for the pore-forming domains (domain I) to insert into the membrane. In the case of bacterial pore-forming toxins active against mammalian cells, this partial denaturation process is stimulated by acidic pH at the cell surface [13]. A similar mechanism could occur with Cry proteins active against lepidopteran insects, although the gut pH is very alkaline; the partial denaturation could still be triggered by a decrease in pH at the cell surface. The pH optimum for aminopeptidase N in lepidopteran larvae (8.0; [46]) is at least 2 pH units less than bulk gut content pH (>10), suggesting that a decrease in pH occurs near the cell surface. The involvement of lipid rafts, microdomains which are less fluid than the membrane as a whole, in pore formation has been suggested [47]. However, membrane-anchored proteins are selectively associated with these lipid rafts, and it is not clear whether lipid rafts are necessary for pore formation, or whether their involvement is a result of the presence of receptors. The trans-membrane cadherin-like *Bt*-R1 receptor is not associated with lipid rafts.

A current model for pore formation by Cry1A toxins suggests that interaction with two receptors is necessary; an initial binding step with the cadherin-like *Bt*-R1 receptor leads to toxin oligomerization, followed

by interaction of the oligomer with the aminopeptidase N receptor and insertion into the membrane [22, 48]. While this model is plausible, the details of the mechanism of toxicity must differ for different toxins, and a “two-receptor” model should not be assumed to be generally applicable. The gain of function experiments described above show that only one receptor is necessary for toxicity to be shown, and only a few lepidopteran-specific Cry toxins have been shown to interact with cadherin-like proteins [49]. If the major determinant of Cry protein toxicity is the assembly of oligomeric complexes at the surface of cells in the gut epithelium, then this requirement can be met in diverse ways, involving different “receptor” proteins to localize the toxin and promote oligomerization (although the interaction is always likely to involve the most abundant proteins at the cell surface). A “global” diversity of interactions is not inconsistent with specificity when interactions between specific toxins and hosts are considered.

Once the insertion of Cry toxin into the cell membrane leads to pore formation, the gut epithelial cell is unable to maintain its internal solute balance, as the open pore allows free exchange of ions and other small molecules between the gut lumen and the cytoplasm. The cytoplasm of gut cells has markedly different concentrations of ions (including H^+) than the gut lumen; this difference in concentrations is used to drive active transport processes, such as amino acid transport [50]. Free movement of ions thus causes massive disruption to cell physiology, leading to death. The leakage of cell contents also causes proliferation of gut microflora, so that dying insects show massive bacterial infection of collapsing gut tissue. Cry proteins may also produce toxic effects through interference with signaling pathways. Binding of Cry1Ab to the transmembrane *Bt*-R1 receptor has been shown to activate a G-protein-mediated intracellular signaling pathway, resulting in the formation of cAMP by adenylyl cyclase, and activation of protein kinase A [51]. This process led to cytological changes typical of *Bt* toxin activity.

Binary Cry and Vip Toxins

The binary Cry toxins are exemplified by toxins active against corn rootworm [52]. These toxins are only active as a combination of two proteins, designated as

families Cry34 (14 kDa protein) and Cry35 (44 kDa protein). The two proteins are the product of a single operon in the commonly used *Bt* strains. The binary toxin acts on the insect gut epithelium, and leads to swelling and vesicle production from epithelial cells, resulting in the disappearance of microvilli, and extensive disruption of the epithelium. However, it is not clear whether these symptoms are solely a result of open pore formation, or whether other modes of toxicity, such as ADP-ribosylation (see below) are occurring. No structural information on these proteins is available at present. There is evidence that the 44 kDa toxin protein Cry35 is evolutionarily related to an insecticidal toxin from *Bacillus sphaericus* [53]. The *B. sphaericus* toxins have received some attention due to their toxicity toward mosquitoes and other dipteran insects. They also bind to membrane-anchored receptors (α -glucosidase, in the case of the mosquito *Culex pipiens* [54]) and cause disruption of the gut epithelium [55]. However, their detailed mechanism of action is not known. Like Cry34/35, the *B. sphaericus* proteins are binary toxins, although in this case one component does show limited activity in the absence of the other. The designation of the corn rootworm binary *Bt* toxin by the symbol Cry obscures the fact that these toxins have little in common with the three-domain toxins, besides being found in crystalline deposits in *Bt*, and being insecticidal as a result of acting on the insect gut epithelium.

The *Bt* insecticidal Vip1/2 proteins (active against corn rootworm) are also binary toxins with similarity to the *B. sphaericus* toxins [56]. The mechanism of action of Vip1/2 toxins involves ADP-ribosylation by the active component, which disrupts actin polymerization in cellular microfilaments, similar to other bacterial ADP-ribosylating toxins such as botulinum toxin [57]. The inhibition of actin polymerization leads to massive disruption of cellular functions. The Vip1Ac binding component of the binary toxin interacts with membranes to form oligomeric channels, allowing the active component to gain access to the cell cytoplasm [58].

A further class of Vip proteins, Vip3, (active against lepidoptera) has been identified; these protein are single chain toxins which lyse insect gut cells by pore formation in membranes, and have no sequence similarity to Vip1/2 [59, 60]. Vip3 binds to brush border

membrane vesicles prepared from target insect gut epithelial cells, but does not bind to the same receptors as Cry1 and Cry2 proteins [61]. Binding to 80 and 100 kDa membrane proteins is observed in ligand binding experiments [62], but these receptors have not been characterized. These proteins are promising candidates for further development; chimeric toxins containing regions from different Vip3 toxins have been produced and show extended ranges of toxicity toward lepidopteran pests [63].

Cyt Toxins

The cytolytic Cyt toxins, also found in crystalline inclusions in some *Bt* strains, are single polypeptides, of approx. 250 amino acids; the N-terminal region contains α -helices which wrap around a C-terminal β -sheet core in the three-dimensional structure [64]. Pore formation results from insertion of the β -sheet region into membranes [65]. Unlike the three-domain Cry toxins, this membrane insertion is not receptor-mediated [66]; the Cyt toxins insert directly into membranes, and are thus cytolytic to a wide range of cells. Like the three-domain Cry toxins, Cyt toxins are synthesized as inactive protoxins which are activated by proteolysis. Activation involves removal of propeptides from both the N- and C-termini of the protoxin; in the case of Cyt2Aa, 32 aa are removed from the N-terminus and 15 aa from the C-terminus to generate active toxin [67]. This process does not require specific proteinases.

The combination of Cry and Cyt toxins found in crystalline inclusions in some *Bt* strains, specifically in the strains of *Bt* subsp. *israelensis* active against mosquito larvae, is highly effective as a toxin due to synergistic interactions between its components. Not only are the three domain Cry protein components in these crystals more effective toxins in the presence of Cyt proteins, but the Cyt proteins also prevent resistance to Cry proteins from developing when insects are exposed to purified protein preparations under laboratory conditions [68]. This synergistic effect could result from the two types of toxin producing complementary disruption of the insect gut epithelial cell membranes, but evidence has been presented that Cry and Cyt toxins can interact directly. Specifically, Cry11Aa and Cyt1Aa bind strongly to each other, both in solution and in a membrane-bound state, and

binding of Cry11Aa to mosquito gut epithelial cell membranes was enhanced by pretreating the membranes with Cyt1Aa [69]. The interaction with Cyt1Aa takes place through the loop region in Cry11Aa involved in protein–protein interactions with its “normal” receptor (membrane GPI-anchored alkaline phosphatase). Insertion of Cyt1Aa into gut cell membranes, which is not dependent on receptor mediation, thus generates additional “receptors” for Cry11Aa, increasing its toxicity, and preventing resistance developing by mutation of the insect-encoded “receptor.”

Expression of Genes Encoding *Bt* Insecticidal Proteins in Transgenic Plants

Expression of Three-Domain Cry Toxins from Transgenes in the Nuclear Genome

Almost all the insect-resistant transgenic crops currently in use express three-domain Cry proteins from *Bt* as their protective agent. The initial laboratory-based experiments expressed Cry1 toxins in plants to give protection against lepidopteran larvae, and this has remained the main focus of *Bt* gene utilization up to the present day. However, the three-domain Cry proteins pose a number of problems in terms of expression in plants. The technology involved in achieving sufficient levels of accumulation of these proteins to give adequate levels of protection was initially challenging, but developed rapidly, so that within 5 years of the initial reports of engineered resistance, the methodology for gene manipulation was essentially complete. The slower pace of transfer of this technology into major crop species observed subsequently has had much more to do with technical difficulties in plant transformation (particularly regenerating viable plants), than with any problems at the level of gene constructs. The minimum level of Cry protein expression in leaf tissue to give high levels of mortality of sensitive lepidopteran larvae under laboratory conditions is approximately 0.05% of total protein, but to give effective field protection against species which are less sensitive to *Bt* toxins, and to manage resistance to the toxin in pests (see later), levels of expression an order of magnitude higher (i.e., 0.5% of total protein) are desirable.

Engineering genes encoding three-domain Cry proteins for expression in transgenic plants has

been extensively described (the review by Mazier et al. [70], gives a particularly comprehensive survey), but a short summary of the main considerations which had to be taken into account is relevant here. These were:

1. How much of the protein coding sequence should be expressed in plants?
2. Which promoters should be used to drive expression of the Cry protein coding sequence in plants?
3. How should the coding sequence be altered to avoid poor expression?

Protein Coding Sequence The C-terminal part of protoxins for three-domain Cry proteins is variable, and absent in some toxins. Its role in directing the formation of crystalline inclusions in *Bt* sporulation is not required when the proteins are expressed in plants (and might result in disruption of cells unless the protoxin was exported into intracellular spaces). All constructs which result in insecticidal activity have omitted this part of the molecule from the coding sequence expressed in plants. The initial research suggests that a complete protoxin accumulates in plant tissue at levels 10–50-fold less than a protoxin truncated so the C-terminal region is absent [3]. Since removal of the C-terminal region of the protoxin does not result in active toxin being produced, retention of the N-terminal activation peptide ensures that the initial protein product in transgenic plant tissue is not active, and proteolytic activation takes place as normal within the gut of insect herbivores. The coding sequence utilized thus corresponds to the three-domain structure shown in Fig. 2, plus the additional N-terminal propeptide.

Promoters The problems experienced in achieving levels of expression of Cry proteins high enough to confer effective protection meant that the initial use of promoter sequences which only gave low levels of expression, such as those from *Agrobacterium tumefaciens* Ti plasmids, was rapidly superseded by strongly expressed promoters, most of which were based on the Cauliflower Mosaic Virus 35S RNA promoter (CaMV 35S). Constitutive expression of

the Cry protein in all plant tissues does not appear to cause significant problems either in a yield penalty, or deleterious effects due to the accumulated protein. However, tissue-specific promoters have also been used, such as the ribulose-bisphosphate carboxylase small subunit promoter (e.g., [71]) or the phosphoenolpyruvate carboxylase promoter (e.g., [72]), both of which are specific for green tissue. The CaMV 35S promoter was initially considered to be specific for dicots, but further experience showed that it could also be functional in monocots, and, with suitable modification, could be used to direct Cry protein expression (e.g., [73]). However, many researchers have preferred to use promoters derived from constitutively expressed monocot genes in Cry protein expression constructs for use in cereal transformation (e.g., the maize ubiquitin-1 promoter; [74]). Root-expressed promoters have been used in constructs designed to protect cereals against corn rootworm [75].

Considerable research has also been undertaken on the use of promoters whose expression is only induced under specific conditions. The use of wound-induced promoters to direct Cry protein expression has the apparent advantage that production of Cry proteins in transgenic plants is, for the most part, only induced on attack by insect pests. Any potential deleterious effects on phenotype caused by production of the toxin in transgenic plants would therefore be minimized, and toxin residues in plant tissues would be reduced. A wound-inducible maize proteinase inhibitor gene promoter has been used to direct expression of Cry1B in transgenic rice, and has been shown to give effective protection against insect attack (against striped stem borer; [76]). However, the protection afforded by transgene constructs containing wound-inducible promoters is lower than when constitutive promoters are used, both in the laboratory and in the field [77].

While achievement of expression levels of *Bt* toxins sufficient to confer protection in transgenic plants is now considered routine, considerable technical problems may still need to be overcome when specific crop species are considered (e.g., soybean; [78]). These include the construction of the synthetic coding sequence for the toxin, choice of an appropriate promoter for the expression construct, developing

protocols for efficient transformation and regeneration of the plant species, and production of homozygous progeny lines containing the transgene.

Engineering the Coding Sequence to Optimize Expression

The initial experiments in which Cry toxins were produced in transgenic plants showed that only low levels of Cry protein were accumulated, generally of the order of 0.01% of total protein, or less. Levels of Cry proteins were at least one order of magnitude lower than when plant proteins were expressed using similar promoters in expression constructs, leading to the deduction that the Cry protein coding sequence contained features which decreased protein production as a result of posttranscriptional events. Cry protein coding sequences are generally A-T rich compared to plants (coding%GC in *Bacillus thuringiensis*, 36%; in *Arabidopsis thaliana*, 45%; in *Oryza sativa*, 55%; Codon Usage Database, <http://www.kazusa.or.jp/codon/>) and codon usages thus differ significantly. Cry protein genes were reengineered, modifying the nucleotide sequence without altering the encoded amino acid sequence, to change the codon usage to one more appropriate for plants, resulting in either partially or wholly synthetic genes (reviewed by Mazier et al. [70]). Codon optimization for both dicots and monocots has been carried out. Codon-optimized synthetic genes show accumulation levels of Cry proteins of up to 1% of total protein in leaf tissue, which is adequate for complete protection of plants against pest insects [79].

The basis for poor expression of Cry proteins in transgenic plants has received comparatively little attention. Evidence suggests that the major problem is not codon usage, but instability of RNA transcripts [80, 81]. Expression of unmodified Cry protein coding sequences leads to accumulation of short, polyadenylated transcripts resulting from incorrect recognition of polyadenylation addition signal sequences within the protein coding sequence [82]. Specific modification of A-T-rich regions within the coding sequence of Cry1Ac toxin putatively responsible for transcription termination and polyadenylation (both AATAAA signal addition sequences and ATTTA upstream motifs) has been shown to lead to increased protein expression in transgenic tobacco [83]. Changing codon usage to increase GC content has eliminated

these A-T-rich regions in synthetic Cry protein genes, which therefore can produce high levels of stable mRNA.

Expression of Three-Domain Cry Toxins from Transgenes in the Chloroplast Genome

The bacterial origin of the chloroplast is reflected in differences in both the genome composition and organization, and the biochemistry of transcription and translation within the organelle, compared to the nuclear genome and transcription and translation in the nucleus and cytoplasm. The bacterial origin of the genes encoding Cry proteins suggests that expression in the plastid, from transgene constructs introduced into the plastid genome, might result in high levels of protein production. This prediction was confirmed in 1995 with a report showing that incorporation of a construct containing a complete coding sequence for the Cry1Ac protoxin protein and the plastid rRNA operon promoter into the genome of tobacco chloroplasts led to accumulation of Cry1Ac protoxin (approx. 130 kDa – i.e., with the C-terminal crystal-forming region intact) in tobacco leaves to levels of 3–5% of total protein [84]. The high level of Cry protein accumulation meant that transformed plants were effectively protected against attack by several major lepidopteran pests, even beet armyworm (*Spodoptera exigua*), a species relatively insensitive to *Bt* toxins.

Despite this highly promising initial report, expression of Cry proteins via plastid transformation has not been widely adopted, and is not used in the current commercial crops. Reasons for this are difficult to pinpoint; there are significant technical problems in achieving stable transformation of plastids, since all of the copies of the plastid genome in the cell (up to 10,000) must be transformed [85], and plastid transformation has been problematic in species other than tobacco [86]. Nevertheless, methods exist to overcome these problems [87]. Cry1, Cry2, and Cry9 proteins have been expressed in plastids of tobacco [88–91], and Cry1Ab has been expressed in soybean plastids [92], all giving high levels of protection against lepidopteran pests to the resulting plants. Overexpression of the Cry2Aa2 operon is particularly effective in giving broad-spectrum protection against a range of pests.

Commercial introduction of transgenic insect-resistant crops based on plastid transformation is almost certainly feasible, but may as yet be restricted by economic considerations, or concern over long-term stability of the transgene phenotype. The maternal inheritance of plastid-encoded characteristics shown by most plants, which means that pollen cannot disperse the transgene to non-transgenic plant stocks, is a further advantage to the method, which could be used to overcome objections to coexistence of transgenic and “organic” agricultural practices by environmental pressure groups.

Expression of Other Genes Encoding Insecticidal *Bt* Toxins

Gene constructs for expression of other *Bt* toxins follow the same principles as those outlined above for three-domain Cry toxins. For example, corn expressing the binary Cry34/35 toxin (for protection against corn rootworm) was transformed with a construct containing a constitutive promoter (maize ubiquitin-1) and synthetic coding sequences for the 44 and 14 kDa polypeptides [52], giving expression levels of up to 0.9% and 0.2% respectively of total soluble proteins in plant tissues. Details of the constructs used for expressing these, and other *Bt* toxins, are apparently not reported in the scientific literature.

Taking Transgenic Plants Expressing *Bt* Toxins into the Field

Dealing with Pest Resistance to *Bt* Toxins

The development of successful strategies for commercial deployment of “first generation” insect-resistant crops expressing a single three-domain Cry toxin has focused on a single major potential problem, the development of resistance to the insecticidal compound by the targeted pest species. Development of resistance to exogenously applied chemical pesticides has occurred in over 500 insect species [93], and field resistance to *Bt* sprays has been observed in the lepidopteran pest diamondback moth (*Plutella xylostella*). Resistance to *Bt* toxins can be produced in the laboratory within a small number of generations of many pests, showing that resistance alleles are present in pest populations at a nonnegligible level, although resistance to high doses

of specific toxins is only shown in individuals homozygous for the resistance allele. This topic has been ably reviewed in the context of the commercialization of *Bt* crops by [94]. The most common mechanism of resistance to Cry toxins in insects is mutation in a toxin receptor, leading to a failure to bind sufficient levels of toxin for lethal effects to be shown; however, the involvement of more than one “receptor” in current models for three-domain Cry toxin mechanisms of toxicity (see above) implies that multiple genetic loci for resistance in the pest are possible. Other mechanisms, such as altered proteolysis of toxins, have been proposed to account for the resistance to multiple toxins which can be produced in the laboratory.

The practical solution to prevent the development of resistance in pest populations, the “high-dose/refuge” strategy, has been extensively reviewed elsewhere [94]. In its simplest form, this strategy couples transgenic plants that are expressing sufficient levels of a specific toxin to kill all pest insects which are homozygous negative, or heterozygous, for a resistance allele, with a reservoir of untransformed plants which maintain a population of pests which have a normal frequency of resistance alleles. It assumes that the frequency of occurrence of resistance alleles is low ($<10^{-3}$). Surviving pests on the transgenic plants will be almost all homozygous positives for the resistance alleles, but will be few in number due to the low frequency of occurrence of these alleles. The non-transformed plants will produce a large number of pest insects, most of which are homozygous negative for resistance alleles. Provided that transgenic and untransformed plants are not spatially separated, mating between resistant insects selected on transgenic plants will be a rare event, and most progeny will be homozygous negative or heterozygous for resistance alleles, and thus susceptible to the insecticidal activity of the transgenic plants. In this way, both the pest population is suppressed, and any increase in the frequency of resistance alleles in the population is minimized by the continuous “diluting out” effect.

This approach has been almost wholly successful in controlling pest resistance to *Bt* toxins in agricultural use of transgenic crops over 10 years. That it has been so successful may be a result of factors other than those originally considered, since the assumption that *Bt* toxin resistance alleles occur at a very low

frequency in natural populations has been called into question. Although some insect populations show resistance allele frequencies in the 10^{-3} to 10^{-2} range (e.g., tobacco budworm, *Heliothis virescens* in USA; [95]; *Sesamia nonagrioides* in Spain and Greece; [96]), estimates for pink bollworm (*Pectinophora gossypiella*) in Arizona, USA in 1997 were as high as 0.16 [97]. No evidence for selection for resistance was observed, since the frequency of resistance alleles did not increase over a 3-year monitoring period in which transgenic cotton expressing *Bt* toxins was extensively employed. A subsequent follow-up study [98] confirmed that frequencies of resistance alleles in this insect had not increased over an 8-year monitoring period, with values generally $<10^{-2}$, despite almost continuous exposure to Cry1Ac via transgenic cotton. The possibility that resistance alleles in the insect carry a significant fitness penalty is one additional factor that could account for these observations.

The success of the refuge strategy is dependent on farmers sacrificing part of their crop (untransformed plants) to maintain a pest population. This has been successfully enforced in the industrialized agriculture of developed countries, but may be more difficult to ensure when insect-resistant transgenic crops become available to rural farmers. Although greater agricultural diversity may play the same role as the refuge strategy in maintaining a pest population and decreasing selection pressure, emergence of resistance in pests to *Bt* crops has been delayed, not eliminated, and further strategies to manage it will be necessary.

Pests That Are Not Susceptible to *Bt* Toxins

As described above, most of the *Bt* toxins that have been investigated, and introduced into transgenic crops, are active against lepidopteran or coleopteran insect pests. This is partly a result of the practical requirements of agriculture, since these orders include most of the major pests. However, there are significant insect herbivores which remain outside the range of activity of *Bt* toxins that have been expressed in transgenic plants.

Dipteran pests, such as fruit flies and root flies, are serious pests in many crops, and *Bt* toxins active against diptera have been thoroughly investigated. A major problem with introducing protection against these

pests into plants is that *Bt* strains active against dipteran insects usually contain a mixture of toxins, often including both Cry and Cyt proteins (see above). These toxins act synergistically, and individual components are only of low toxicity. Introduction of genes encoding the mixture of toxins found in a typical dipteran-active *Bt* strain into a transgenic plant has yet to be attempted, although it is not beyond the capacity of existing technology.

The major order of insect herbivores outside the range of *Bt* toxins is Hemiptera, which includes aphids, plant- and leafhoppers, whitefly, and other sap-suckers which feed directly on the contents of phloem and/or xylem vessels, predominantly sucrose and free amino acids. These insects are important pests and virus vectors. No *Bt* toxins with activity against them have been found. The reason for this is not clear; receptors similar to those in other insect orders are present in these insects [99], but generally they contain very low levels of digestive proteolytic activity, as a result of ingesting nitrogen in the form of amino acids rather than protein. This lack of digestive proteolytic activity may interfere with activation of *Bt* toxins, and prevent enough activated toxin to have effects on the insect being present in the gut.

Why Haven't Plants Evolved Their Own *Bt*?

Despite the problems encountered in managing resistance of pests to *Bt* toxins, transgenic plants expressing these insecticidal proteins have proved their value in the field. However, the necessity for resistance management suggests that this solution to defense of plants against insect herbivores may not be viable on an evolutionary timescale. Endogenous expression of *Bt* toxins is not a "natural" method of defense against herbivores, since plants do not produce similar insecticidal proteins themselves. This failure on the part of plants to exploit a viable strategy for protection seems puzzling, and the obvious explanation, that plants lack the capacity to produce *Bt* toxin-like proteins, is not correct. Since introduction of suitably modified *Bt* genes gives adequate levels of protein expression for protection, there is no reason why plants could not have evolved a similar capacity. As discussed in the following section, plants have evolved a diverse array of defensive mechanisms, but make little use of

proteins which are highly toxic to insect herbivores. Possibly, this is due to the relative ease with which insects can develop resistance to protein toxins which exert a very strong selection pressure on the population; although alternative hypotheses, such as the balance between investing plant resources into defense versus growth not favoring this strategy, or practical difficulties for a sessile organism in delivering toxins, should also be considered. Unfortunately, the experiments which would enable this issue to be investigated, namely, an evaluation of the "fitness" of *Bt*-expressing plants in a natural ecosystem in competition with varieties relying on endogenous defenses, and the persistence of *Bt* genes in a natural population, are unlikely to be carried out in the near future, due to obvious regulatory issues.

Whatever the reason for plants "in the wild" not using defensive proteins similar to *Bt* toxins, there is no reason to suppose that transgenic plants with engineered insect resistance will not continue to be useful in the artificial growing conditions of agriculture. Manipulation of crop plants by conventional breeding has successfully introduced characteristics such as large seed size, which were not present, and would not be viable, in their wild progenitors. Characteristics introduced into cultivated plants by plant genetic engineering do result from a process that is fundamentally different from selection, but both conventional breeding and genetic engineering are aiming for the same end results, agriculturally desirable phenotypes. Their products should be evaluated by similar criteria.

Developments to "First-Generation" Crops Expressing *Bt* Toxins

Plants Expressing Multiple Toxins ("Pyramiding")

The specificity of a single Cry toxin toward specific target pests can be a problem in the field where a secondary, minor pest species can replace the primary pest and cause serious damage to crops. An obvious method to counter this problem is to add or introduce a second *Bt* *cry* gene into the crop to extend the range of pests against which protection is afforded. The availability of a wide range of gene constructs encoding Cry toxins has made this a realistic possibility, with crossing singly transformed lines, or repeated transformation,

or transformation with a construct containing two genes as alternative methods for introducing the genes into one line. Monsanto's Bollgard transgenic cotton was improved by introducing a second *Bt* gene as early as 1999. Laboratory trials showed that cotton plants expressing both Cry1Ac and Cry2Ab proteins were more toxic to bollworms (*Helicoverpa zea*) and two species of armyworms (*S. frugiperda* and *S. exigua*) than cotton expressing Cry1Ac alone, even though doses in this trial were sublethal [100]. Subsequent evaluations in greenhouse and field trials [101] confirmed the superior insect resistance of plants expressing both toxins.

A further potential advantage of transgenic plants expressing two Cry proteins with differing specificities, that target different receptors in the insect, is in preventing the appearance of resistance in the pest, since multiple mutations are required to produce the loss of sensitivity to the toxins. This hypothesis was confirmed directly in work reported by [102], in which transgenic broccoli plants expressing either Cry1Ac, or Cry1C, or both proteins were produced. Plants were exposed to a population of diamondback moth (*P. xylostella*) which carried *Bt* resistance genes at a relatively low frequency in an extended greenhouse experiment, and results showed that selection over 24 generations led to a significant delay in the appearance of resistance in insects exposed to the pyramided two-gene plants. The success of these experiments has led to suggestions that the refuge approach to resistance management may be redundant for crops expressing multiple toxins [103]. However, some care is needed in the selection of genes in relation to potential pests, as resistance to multiple toxins has been observed in several cases. For example, a strain of the lepidopteran cotton pest *H. virescens* which has simultaneous resistance to Cry1Ac and Cry2Aa has been identified, in which the genetic bases of resistance to each toxin are different [104].

Many subsequent programs which have aimed to produce insect-resistant crops expressing *Bt* toxins have adopted the two-gene approach to broaden and improve protection against diverse pests, and to prevent resistance developing in insects (e.g., [105]). Although engineering to produce combinations of different three-domain Cry toxins is the most common approach, other potential resistance genes have been

included also, such as those encoding Vip proteins [106], or even proteinase inhibitors (e.g., cowpea trypsin inhibitor; [107]). The "pyramiding" or "stacking" of resistance transgenes has been enthusiastically adopted by commercial organizations, and the recent announcement of a transgenic maize variety containing eight different transgenes by Monsanto and Dow Agrosciences [108] exemplifies this trend. This variety contains insect-resistance genes derived from both companies' research programs, active against corn rootworm and lepidopteran pests (Herculex RW = Cry34Ab1 + Cry35Ab1, Herculex I = Cry1F; YieldGard VT Rootworm/RR2 = modified Cry3Bb1, YieldGard VT PRO = Cry1A.105 + Cry2Ab2), as well as two herbicide tolerance genes (giving resistance against glyphosate and glufosinate-ammonium), and is intended to be a "one-stop" solution to pest and weed problems.

Domain Exchange in Three-Domain Cry Toxins

The separate roles played by the different domains in the process of interaction of three-domain *Bt* toxins with their receptors, and their structural independence, suggested to investigators that hybrid toxins, in which domains from different naturally occurring toxins were grafted together, would be likely to be active, and could show novel specificities in their activity toward insects. This process can be made to occur in vivo in *Bacillus thuringiensis*, using a site-specific recombination vector [109], or can be carried out in vitro using conventional molecular biology techniques, followed by expression in a microbial host. Transfer of domain III between different Cry1 proteins led to identification of this domain as conferring primary specificity to different lepidopteran species, and the generation of hybrids with broader specificity than naturally occurring toxins [110]. Subsequent work generated a Cry1Ab-Cry1C hybrid, which was highly toxic to *S. exigua*, an insect resistant to Cry1A toxins [111], and identified Cry1Ca domain III as sufficient to confer toxicity toward *Spodoptera* in a variety of hybrids [112]. In contrast to the results obtained when exchanging domain III, exchange of domain I between different Cry1 toxins did not yield biologically active proteins [113].

A measure of the potential for improvement in "natural" *Bt* toxins is shown by experiments reported by [114], in which a hybrid Cry protein, constructed by

fusing domains I and III from Cry1Ba with domain II of Cry1Ia, was expressed in transgenic potato. Plants expressing the hybrid toxin at levels up to 0.3% of total soluble protein were produced, and not only showed resistance to the lepidopteran pest potato tuber moth (*Phthorimaea operculella*), but also had a high level of resistance to Colorado potato beetle. The “parental” Cry proteins have high toxicity towards lepidopterans, but only very limited toxicity towards coleopterans such as the potato beetle. The hybrid has effectively created a novel toxicity, which is suggested to be based on interaction with a novel receptor.

Mutagenesis of Three-Domain Cry Toxins

Modification of *Bt* toxins by site-directed mutagenesis to increase toxicity towards target pests has been employed as an alternative to the “domain swap” approach. Most mutagenesis experiments on *Bt* toxins have been carried out to explore structure-function relationships in these proteins (see above; reviewed by Dean et al. [115]), but the accumulated knowledge of which parts of the protein determine specificity of interactions with receptors in the insect have been exploited to produce variants with increased activity toward target pests.

The key role of domain II in three-domain Cry proteins in mediating interactions with insect receptors was shown by a mutagenesis experiment in which altering amino acid residues in the loop regions in this domain of Cry1Ab increased its toxicity toward larvae of gypsy moth (*Lymantria dispar*) by up to 40-fold, with a corresponding increase in binding affinity to brush border membrane vesicles [116]. These results were based on expression of the recombinant protein in microbial hosts. A similar strategy was used to increase the toxicity of Cry3A protein toward target coleopteran pests [117], and of Cry4Ba toxin [118, 119] and Cry19Aa toxin [120] toward mosquito larvae. The level to which rational design of toxins is possible is shown by the engineering of toxicity toward mosquito larvae into the lepidopteran-specific toxin Cry1Aa [121]. Alternatively, a directed evolution system based on phage display technology for producing toxins with improved binding to a receptor, and thus increased toxicity, has been described [122]. Mutagenesis of domain I has also been attempted, with claims that

alteration of alpha helix 7 in Cry1Ac to resemble the corresponding helix in diphtheria toxin led to increased toxicity toward cotton bollworm (*Helicoverpa armigera*) larvae [123].

The impressive achievements of toxin engineering at the level of recombinant proteins, have led to the technology being used for gene constructs designed for expression in transgenic plants, although toxins with unmodified amino acid sequences continue to be widely used (largely as they give adequate protection). One example where toxin engineering has been successfully carried out is the current commercial transgenic corn variety with resistance to corn rootworm, MON863, which expresses a modified version of the *Bt* Cry3Bb1 toxin [75]. Unmodified Cry3Bb1 is active against a number of coleopteran species, including Colorado potato beetle and corn rootworm [124], but toxicity toward western corn rootworm (*Diabrotica virgifera virgifera*) was not sufficient to give adequate protection at levels of expression achievable in corn. Modifications to the amino acid sequence increased the toxicity of the protein toward corn rootworm approximately eightfold. The nature of the modifications has not been described in the scientific literature, and is only available through reference to a series of patents (see [75]).

Fusions

As a logical extension to the transformation of plants with separate gene constructs encoding two Cry proteins, some workers have chosen to produce a single construct containing a single translationally fused coding sequence encoding both proteins. This approach has been successfully demonstrated by producing a Cry1Ab-Cry1B translational fusion protein in transgenic maize [125], although there is no apparent advantage over simpler methods for introducing two genes. The Cry1Ab-Cry1B fusion protein has also been expressed in transgenic rice [126], which was fully resistant to yellow stem borer (*Scirpophaga incertulas*).

A more interesting possibility is the introduction of extra functionality into Cry toxins by addition of sequences from other proteins which could lead to binding interactions with more potential receptors in the insect gut, extending the range of toxicity and hindering development of resistance. In work reported

by Mehlo et al. [127], the galactose-binding lectin domain (B-chain) from the ribosome-inactivating protein ricin was fused C-terminally to domain III of Cry1Ac, producing a Cry1Ac-ricin B-chain fusion protein. The fusion protein thus has the ability to bind to galactose residues in side chains of glycoproteins or glycolipids in the insect gut epithelium, as well as *N*-acetyl galactosamine residues which are bound by domain III. The fusion protein was expressed in transgenic maize and rice plants, and was shown to afford a high level of protection to larvae of stemborers (*Chilo suppressalis*) and leaf armyworm (*Spodoptera littoralis*), whereas plants expressing the unmodified Cry1Ac were susceptible to both insects. The transgenic maize plants were also resistant to a homopteran plant pest, the leafhopper *Cicadulina mbila*, although it is possible that this was an effect of the lectin domain in the fusion (see later section Lectins), since *Bt* toxins are not effective against homopteran insects.

The engineering of extended binding properties into three-domain Cry proteins to increase the range of toxicity toward insect pests is clearly possible, but needs to be approached with some caution. There is a risk that the extended range of activity will include mammalian toxicity, which would negate one of the major advantages of these insecticidal proteins.

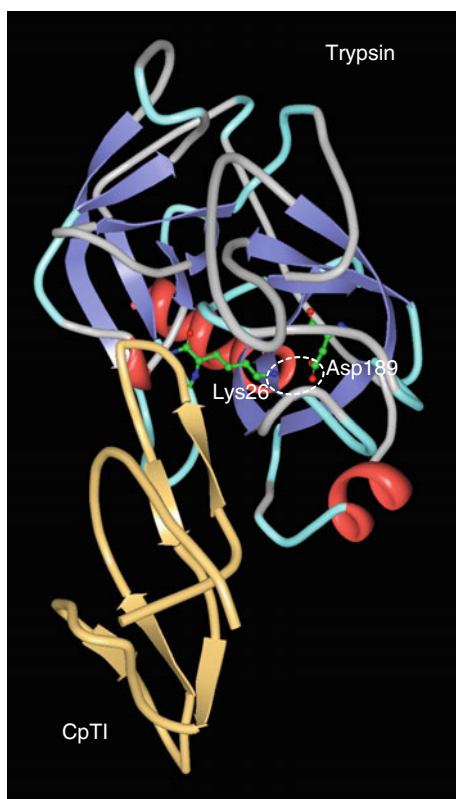
Exploitation of Endogenous Plant Defensive Mechanisms Against Insect Herbivores

Plants have a range of endogenous mechanisms to defend themselves against insect herbivores, and use both static defense mechanisms based on the accumulation of pre-synthesized insecticidal compounds, and active defense mechanisms in which gene expression is induced as a result of insect damage (response to wounding, and responses to insect secretions), leading to the synthesis of insecticidal compounds [128]. Conventional breeding has sought to exploit endogenous insecticidal genes within a plant species, but the use of transgenic technology allows defensive compounds and mechanisms to be transferred between species, or allows the control of existing defensive systems to be altered to improve their effectiveness. The molecular biology involved in transfer of genes between plant species is technically straightforward, and does not involve the kind of reengineering necessary to make

bacterial genes suitable for use in plants. This approach to increasing insect resistance in transgenic plants has almost as long a history as engineering for *Bt* Cry toxin expression, but to date has not resulted in a commercial product, or widescale adoption in agriculture. Some of the reasons for the lack of practical outcomes for this strategy will be discussed below.

Proteinase Inhibitors

Protein proteinase inhibitors (PIs) are ubiquitous in plant species. They are major components of both “static” and “active” defense in that they are accumulated in specific tissues (“static” defense), and are the major end-product in the induced response to wounding (“active” defense). They are generally small proteins, ranging in size from 4 to 25 kDa, with many different sequence families having been identified. They form tightly bound complexes with their target proteinases, which usually involve a “loop” on the inhibitor fitting into the enzyme active site (Fig. 4), blocking the site, and inactivating the enzyme. The observation that most of these inhibitors were active against digestive serine proteinases from higher animals, and not endogenous plant proteinases (where serine proteinases are comparatively rare, and not involved in protein digestion) suggested that they were defensive compounds, and bioassays in which purified PIs were fed in artificial diet confirmed that an antimetabolic effect was exerted on insect herbivores which relied on protein digestion for nitrogen supply, shown as a slower growth rate, retarded development, and increased mortality (reviewed by Garcia-Olmedo et al. and Ryan [130, 131]). Besides a direct effect on digestion of ingested proteins, PIs cause a loss of nitrogen to the insect by preventing the reabsorption of nitrogen used to produce digestive proteinases, which are normally (self)-degraded in the gut rather than excreted. The role of these proteins in induced defense against insects was shown by blocking the normal wounding response in transgenic tobacco plants by suppression of expression of the prosystemin gene, which produces the peptide hormone systemin, using antisense RNA. The transformed plants were unable to synthesize wound-induced PIs and were significantly more susceptible to herbivory by lepidopteran larvae [132]. The importance of the wounding response to



Genetic Engineering of Crops for Insect Resistance.
Figure 4

Structure of a complex between a typical plant protein proteinase inhibitor (PI) and a target proteinase (from PDB 2g81; [129]). Structure shown in backbone representation is the complex between beta-trypsin (*top*, secondary structure color-coded in red and blue) and a Bowman-Birk PI from cowpea (*Vigna unguiculata*; *bottom*, gold). This inhibitor ("CpTI") has been expressed in transgenic plants to give partial resistance to lepidopteran larvae. The side chains responsible for the specificity-defining ion-pair interaction (dotted ellipse) are shown in ball-and-stick representation; they are Asp189 (S1') in the substrate binding pocket of the enzyme, and Lys26 (S1) on the active site loop of the inhibitor. Other interactions take place across the contact surface between inhibitor and enzyme to form a tightly bound complex

plant defense in natural ecosystems has been extensively studied by Baldwin's group (reviewed in [133]); this outstanding body of work has established a synthesis of responses in the plant under attack, responses

in neighboring plants, and responses of natural enemies of insect herbivores, with communication via volatile signals produced by the plant under attack.

A seed-expressed Bowman-Birk-type serine proteinase inhibitor from cowpea, which contained two inhibitory sites active against bovine trypsin (CpTI) was the first plant PI to be produced in another species [134], using a gene construct containing a CaMV 35S promoter. The resulting transgenic tobacco plants expressed CpTI at up to 1.0% of total soluble protein, and decreased growth and survival of tobacco budworm (*H. virescens*) by up to 50%, with similar effects on other lepidopteran larvae. Subsequent experiments carried out with wound-induced PIs showed that these also had similar effects when constitutively expressed in transgenic plants; for example, the tomato inhibitor II gene, when expressed in tobacco, was also shown to confer insect resistance [135], as did potato PI-II [136]. Both CpTI and PI-II were subsequently expressed in rice, where partial protection against stem borers was observed [137, 138]. The constitutive expression of foreign PIs could be mimicked in transgenic tomato plants by constitutive expression of the prosystemin gene (see above) leading to constitutive expression of wound-induced tomato PIs [139]. Tobacco plants modified in this manner show partial resistance to insect herbivores similar to that produced by expressing foreign PIs [140].

The problem with this strategy for producing insect-resistant plants soon became obvious; in contrast to the expression of Cry proteins, which, when optimized, routinely gave transgenic plants virtually complete protection against susceptible pests (mortality 100%, damage minimal) expression of PIs only produced partial resistance. Investigation of the digestive biochemistry showed that exposure to PIs in the diets of lepidopteran and coleopteran herbivores resulted in the appearance of proteinase activities which were insensitive to the inhibitor(s) present [141, 142], or were able to degrade the ingested PIs [143]. These insects contain large families of genes encoding dietary proteinases, whose expression could be up- or downregulated by dietary inhibitors [144]. In effect, these insect herbivores were preadapted to be partially resistant to dietary PIs, as a result of similar or identical compounds being present routinely in their diet. Although expression of resistance to PIs in herbivorous

insects has a fitness penalty, shown by reduced growth on diets to which inhibitors are added, or on plants which are expressing foreign PIs, or over-expressing endogenous PIs (see above), this is not sufficient to cause mortality at a level which affords more than partial protection. In some cases, low levels of expression of a foreign PI in transgenic plants can actually result in improved insect performance, as when tobacco and *Arabidopsis* plants expressing mustard trypsin inhibitor 2 were exposed to larvae of cotton worm (*S. littoralis*; [145]).

A number of investigators have attempted to select PIs for expression in transgenic plants which are optimally active against the dietary proteinases present in specific insect pests. Attempts to develop inhibitors active against specific lepidopteran digestive serine proteinases induced by dietary PIs have not been successful. On the other hand, not all pest insects rely on serine proteinases for digestion. Many herbivorous coleopteran larvae utilize cysteine proteinases, rather than serine proteinases, as their major digestive endoproteinases, and these proteinases can be inhibited by cystatins, a family of proteins present in all kingdoms of organisms. Enzyme assays in vitro were used to characterize digestive proteinases of a coleopteran pest, *Chrysomela tremulae*, as cysteine proteinases, and to show that a cystatin from rice, oryzacystatin, was an effective inhibitor. Transgenic poplar seedlings expressing oryzacystatin were produced, and leaves from these plants were shown to be toxic to larvae of the pest [146]. This promising result does not seem to have been followed up. Expression of oryzacystatin in transgenic potato only gave partial protection against larvae of Colorado potato beetle [147], suggesting preadaptation in this pest, which is known to employ a diverse range of digestive proteinases. In an attempt to use proteinase inhibitors which insects would not be preadapted to, synthetic multi-domain cysteine proteinase inhibitors based on domains found in animal and plant sources (kininogen, stefin, cystatin C, potato cystatin, and equistatin) were assembled and expressed in transgenic potato; the plants were deterrent to thrips, and gave partial resistance in greenhouse trials, but complete protection was not observed [148, 149]. Attempts to express the sea anemone cysteine/aspartic proteinase inhibitor equistatin itself in transgenic potato did not give significant levels of resistance to Colorado potato beetle, due to degradation of

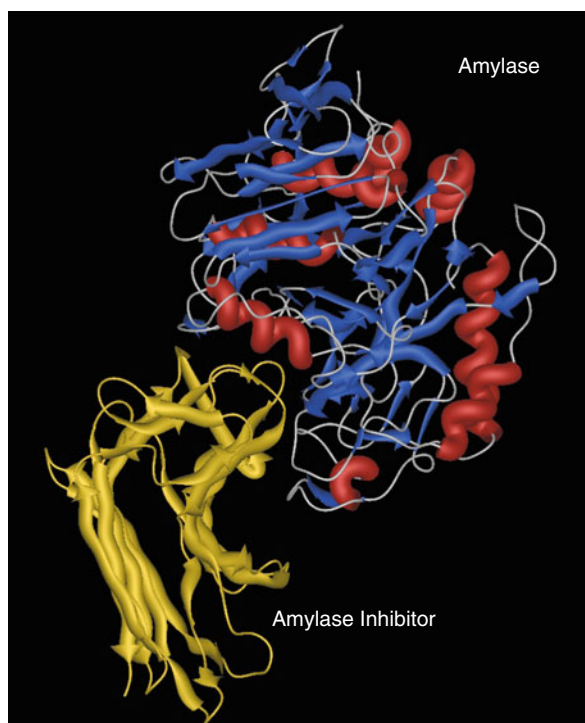
the inhibitor in the plant [150]. Multiple proteinase inhibitors (potato PI-II and PCI) active against two families of proteinases, serine proteinases and carboxypeptidases, have been expressed in transgenic tomato plants [151], but still only afforded partial protection against lepidopteran larvae due to adaptive mechanisms present in the insects.

In conclusion, the expression of suitable PIs in transgenic plants can give protection against lepidopteran and coleopteran pests, but has not been able to produce results comparable with those achieved by use of *Bt* toxins.

Amylase Inhibitors

The widespread occurrence of protein inhibitors of mammalian amylases in plants has become accepted as another defensive mechanism against herbivores (reviewed by Franco et al. [152]). Like proteinase inhibitors, these are generally small proteins, resistant to proteolysis, ranging in size from approx. 8–30 kDa. Although they are also active against insect amylases, it is not clear to what extent these proteins contribute to insect resistance in most cases, since the relatively low nitrogen content of plant tissues compared to insects means that most herbivorous insects are nitrogen limited, not carbon limited, and starch digestion is unlikely to be a limiting factor in growth. However, in the case of coleopteran herbivores whose larvae attack seeds specifically, such as seed weevils (bruchids), there is good evidence for α -amylase inhibitors from legume seeds being highly insecticidal [153], and in being causative factors in the resistance of specific varieties of legumes to bruchids [154]. These proteins belong to a different sequence family than the more common types of α -amylase inhibitors found in cereals, and are similar to legume lectins in sequence [155].

Like proteinase inhibitors, amylase inhibitors form tightly bound complexes with their target amylase (Fig. 5), although the same interaction of a loop on the inhibitor with the active site of the enzyme is not possible, since the enzyme substrate is a polysaccharide, not a polypeptide. The mechanism of toxicity clearly involves inhibition of starch digestion, since bruchid larvae exposed to the α -amylase inhibitor from French



Genetic Engineering of Crops for Insect Resistance.

Figure 5

Structure of a complex between a plant protein α -amylase inhibitor and an insect amylase enzyme (from PDB 1viw; [156]). Structures shown in backbone representation; α -amylase from larvae of the coleopteran storage pest *Tenebrio molitor* (yellow mealworm beetle) is shown *top right*, in red and blue (secondary structure color coding); the α -amylase inhibitor from *Phaseolus vulgaris* (French bean) is shown in gold *bottom left*. This inhibitor has been expressed in several transgenic legume species to give resistance to coleopteran pests. The inhibitor shows the typical “all β -sheet” structure of the legume lectin family of proteins. Interaction between the binding loop of the protein and the starch-binding site of the enzyme occurs across the contact surface, sterically blocking access by polysaccharides to the active site

bean (*Phaseolus vulgaris*) show induction of amylase enzymes [157], although other mechanisms of toxicity may also be present, since these proteins can cause 100% mortality in susceptible insect species at levels of <1.0% of total protein. Alternatively, these highly specialized herbivores may lack the adaptive mechanisms to plant defensive proteins shown by species that

feed on a wide range of plant foodstuffs [128]. High levels of toxicity toward insects have not been observed in general with amylase inhibitors. For example, α -amylase inhibitors are not strongly toxic to lepidopteran larvae, where the alkaline environment of the gut may interfere with the formation of inhibitor-enzyme complexes. The α -amylase inhibitor from French bean is inactivated by high pH.

The isolation of a lectin-like α -amylase inhibitor gene from *P. vulgaris* [155] stimulated research in this area, and in a ground-breaking series of experiments, this gene was assembled into a construct with a strong seed-specific promoter (from the *P. vulgaris* seed lectin gene), and expressed in seeds of transgenic garden pea. The resulting seeds contained up to 3% of the foreign protein, and were highly resistant to larvae of cowpea and Azuki bean weevils [158], which do not normally attack garden peas in the field, but are stored product pests, and to larvae of the pea weevil *Bruchus pisorum* [159], which is a field pest of garden pea. In all cases larval development from eggs laid on seeds was halted at a very early stage, and damage to the crop was minimal. Subsequent experiments showed that transgenic azuki beans could also be protected against bruchid storage pests [160], and that transgenic garden pea was protected against pea weevil under field conditions [161]. The success of this strategy led to hopes that the *Phaseolus* α -amylase inhibitor gene could be incorporated into a range of crops, particularly other grain legumes such as lentils, mungbean, groundnuts, and chickpeas to give protection against a variety of bruchids. Technical problems with transformation of some of these crop species have delayed this goal being achieved, but transgenic chickpeas expressing high levels of the *Phaseolus* α -amylase inhibitor have been successfully produced [162].

Despite the success of this strategy, full agricultural deployment of transgenic crops expressing the *Phaseolus* α -amylase inhibitor gene has not taken place. Commercial reasons have played a major part in preventing widescale adoption, but safety concerns have also arisen. The protein product of the *Phaseolus* α -amylase inhibitor gene expressed in pea shows minor structural differences to the native product (i.e., expressed in *P. vulgaris*) as a result of differences in posttranslational processing (differences in the extent of glycosylation, and in minor components

resulting from proteolysis). Whereas consumption of the native form of the *Phaseolus* α -amylase inhibitor by mice did not result in immunological responses, consumption of transgenic peas expressing this protein led to the presence of circulating antibodies directed against it, and systemic immunological responses including inflammatory responses (i.e., allergic responses) to inhaled or injected protein [163]. In contrast to some earlier work claiming that consumption of transgenic plant material was harmful, this study has been published in a fully peer-reviewed journal and the quality of the research has not been disputed. Further research will be necessary to identify, and remove, the cause of this increased antigenicity. An additional potential drawback was revealed by feeding trials of transgenic peas expressing *Phaseolus* α -amylase inhibitor with pigs and chickens. These trials did not show immunological effects on animal health, but did show that starch utilization by the animals was significantly decreased due to the presence of the inhibitor in the transgenic peas when compared to non-transgenic peas, consistent with the effect of the protein on higher animal amylases [164, 165]. This factor would limit the utility of transgenic peas as animal feed.

Lectins

Lectins, or carbohydrate-binding proteins, occur throughout the plant kingdom, and in many species are accumulated in plant tissues as defensive proteins, being particularly abundant in seeds and other storage tissues, where they can account for up to 1% or more of total protein (reviewed by Peumans and van Damme and van Damme et al. [166, 167]). They are multimeric proteins containing polypeptides which range from 10 to 35 kDa in size. The insecticidal activity of lectins was first observed in assays with larvae of coleopteran species (e.g., LE QA done [168, 169]), where retardation of development, and in some cases, mortality, was observed when lectins were incorporated into diets at 1–5% of total protein. Lectins have only relatively low antimetabolic effects on lepidopteran larvae when fed in diet [170], possibly as a result of high gut pH inactivating the carbohydrate-binding activity. The mechanism of toxicity of these proteins remains obscure, but is dependent on carbohydrate binding.

Although transgenic tobacco and potato plants expressing lectins from garden pea [171] and snowdrop [172] have been produced by standard transformation techniques, and have been shown to confer partial resistance to lepidopteran larvae (>50% reductions in plant damage, with increased larval mortality and decreased growth), the availability of better insecticidal genes specific for these pests has directed this approach toward different targets. Homopteran plant pests, which are not affected by known *Bt* toxins, were shown to be susceptible to lectin toxicity when the proteins were delivered via artificial diet [173]. Susceptibility varied between species, and between lectins, but LC_{50} values as low as 6 μ M have been estimated (for snowdrop lectin fed to rice brown planthopper (*Nilaparvata lugens*); [174]). Expression of the mannose-specific snowdrop lectin (*Galanthus nivalis* agglutinin; GNA) in transgenic rice plants was carried out, using both a phloem-specific (rice sucrose synthase) and a constitutive (maize ubiquitin-1) promoter [175]. The resulting plants were partially resistant to rice brown planthopper, with reductions of up to 50% in survival of immature insects to adulthood, and reduced development and fertility of survivors. Results were confirmed by independent transformations of indica rice varieties [176]. GNA-expressing rice was also resistant to other homopteran plant pests, such as green leafhopper (*Nephotettix virescens*; [177]) and whitebacked planthopper (*Sogatella furcifera*; [178]). Plants expressing both GNA and Cry1Ac were protected against both brown planthopper and striped stem borers (*C. suppressalis*), but no synergistic effects between the two insecticidal proteins was observed [179]. Further progress on this research has been limited, due to concerns about possible adverse consequences to higher animals of ingesting snowdrop lectin. While earlier data must be regarded as unreliable, a recent study found that no adverse effects of consumption of transgenic rice expressing GNA by rats, although significant differences in some parameters to a control group were observed [180]. GNA expression has also been engineered into potato [181] and maize [182], to give partial resistance to peach-potato aphid (*Myzus persicae*) and corn leaf aphid (*Rhopalosiphum maidis*), respectively. However, these insects are insensitive to lectin toxicity, and only marginal effects on fecundity were observed.

Introduction of foreign lectin genes into plants has become established as a potential method for engineering insect resistance, although with the lectins tested at best only partial protection against homopteran pests is conferred, and some species are relatively insensitive to the effects of lectins. As is the case with PIs, it is likely that plant pests are preadapted to the presence of lectins as defensive compounds, and are able to tolerate the toxic effects to varying degrees, although responses induced in insects by ingested lectins have not been characterized. Attempts have been made to select lectins which are the most effective toxins against target insect pests; a mannose-specific lectin expressed specifically in garlic leaves (ASA-L) was observed to show a high level of toxicity toward homopteran pests [183]. A gene encoding this lectin has been engineered into a variety of transgenic plant species, including tobacco [184] and Indian mustard [185], in both cases producing partial resistance to aphid species, with reduced survival and fecundity. Expression of this lectin in transgenic rice using constitutive [186] or phloem-specific promoters [187] gave protection against homopteran pests comparable to, or slightly better than, earlier transformations using gene constructs encoding GNA. The transgenic rice plants expressing ASA-L were shown to decrease transmission of Rice Tungro Virus by its insect vector (green leafhopper), presumably by causing decreased feeding by the pest [188].

Despite these encouraging results, widescale adoption of transgenic crops expressing lectins will probably not occur unless a major commercial company is able to gain exclusive marketing rights, and invests in pushing the transgenic varieties through the regulatory process. This is unlikely to happen, as the technology is not readily protectable by patenting.

Oxidative Enzymes

Induction of polyphenol oxidase (PPO) synthesis is one of the end-results of the plant wounding response [189], and it would seem reasonable to suppose that increased levels of this enzyme would lead to enhanced resistance to insect attack. PPO activity leads to tissue browning, which has been correlated with enhanced insect resistance. The oxidative cross-linking of tannins to proteins catalyzed by PPO decreases protein digestibility, and limits nitrogen availability [190]. However,

there is little or no evidence that PPO levels are correlated with insect resistance (e.g., [191]). High-level, constitutive over expression of a poplar PPO gene in transgenic poplar seedlings led to levels of PPO up to 50x higher than normal in plant tissues [192], but these plants had only marginal effects on larvae of the lepidopteran insect pest forest tent caterpillar (*Malacosoma disstria*). No feeding deterrence was observed, and there was no effect on larval growth or survival except under conditions where larval survival was poor on controls. PPO activity was detected in insect gut and frass, so the negative results were not due to enzyme inactivation. The conclusion that herbivorous insects are preadapted to be able to deal with PPO activity, as a result of exposure to the wounding response on an evolutionary timescale (in a similar manner to preadaptation to PIs – see above) is difficult to avoid.

Peroxidase activity is also induced when plants are stressed, or attacked by pathogens, as part of a lignification response, and several attempts have been made to over-express peroxidases in transgenic plants to enhance insect resistance, despite a lack of clear-cut evidence that peroxidase activity in plant tissues is toxic to insect herbivores. Initial results using tobacco as the host plant, with over-expression of tobacco anionic peroxidase, showed only marginal effects [193], although limited broad-range protection against a variety of pests was observed in the field [194]. The limited protection afforded by this technique argues against further development.

Other Plant Proteins

Ribosome inactivating proteins (RIPs) and chitinases have also been viewed as defensive proteins in plants, although it is not clear that they are part of defense against insect herbivores. Both types of proteins have been expressed in transgenic plants, with variable results in conferring insect resistance. Expression of a maize RIP in transgenic tobacco resulted in very low levels of protection against corn earworm (*H. zea*), which were barely statistically significant [195]. Plant chitinases in general show low toxicity toward insects, but a poplar chitinase, designated WIN6, was selected on the basis that its expression was induced by insect attack. Expression of WIN6 in transgenic tomato plants

led to partial protection against larvae of Colorado potato beetle, with retardation of larval development observed [196]. Expression of the chitin-degrading enzyme *N*-acetylhexosaminidase from *Arabidopsis* in various transgenic plant tissues also gave some protective effects against lepidopteran larvae [197], but it is difficult to see what advantages over other strategies this approach could give. Orally ingested insect chitinases are strongly toxic to lepidopteran larvae (e.g., [198]). However, expression in transgenic plants gave only partial protection against insect herbivores [199], or, in one case, increased susceptibility to attack [200]. Expression of chitinase A from baculovirus AcMNPV in transgenic tobacco gave similar results, with only small effects on lepidopteran larvae and aphids [201].

Engineering Secondary Metabolism for Plant Defense

Compounds synthesized as the end-products of secondary metabolism play major roles in both constitutive and induced defense against insect herbivores in many plant species (reviewed by Wittstock and Gershenon [202]). The idea that these compounds could be used as insecticides has been a part of agriculture for thousands of years, and has been exploited successfully by synthetic chemistry in the production of classes of insecticides such as pyrethroids, based on terpenoid esters produced by flowers of pyrethrums (*Chrysanthemums*). Although the concept of synthesizing a foreign, insecticidal secondary metabolite in a transgenic plant developed concurrently with plant transformation technology, the biosynthesis of most secondary compounds was poorly understood, and the necessity of cloning and introducing a series of genes expressing biosynthetic enzymes to produce a secondary metabolite was considered beyond the techniques available at the time. Anticipation of problems in ensuring controlled co-expression of a series of biosynthetic genes has proved to be over-pessimistic, and plants containing multiple expressing transgenes have been produced without difficulty.

The explosion of knowledge brought about by large-scale cDNA sequencing programs and the *Arabidopsis* genome program has resulted in a much better understanding of secondary metabolism, with

many biosynthetic pathways now reasonably well understood, and clones encoding biosynthetic enzymes available. The first successful demonstration that a foreign secondary compound could confer insect resistance in a transgenic plant [203] exploited a biosynthetic pathway for cyanogenic glycosides. The cereal *Sorghum bicolor* produces a cyanogenic glycoside, dhurrin, by a biosynthetic pathway starting from the amino acid tyrosine, a product of primary metabolism. Two oxidation reactions catalyzed by cytochrome P450 oxidases generate *p*-hydroxymandelonitrile, which is then glycosylated by a UDP-glycosyltransferase to form dhurrin. The three sorghum enzymes responsible were cloned and assembled into expression constructs using constitutive (CaMV 35S) promoters [204], and *Arabidopsis* plants were successively transformed with a construct containing both P450 oxidase sequences, and the glycosyl transferase sequence. All the enzymes were localized correctly (to endoplasmic reticulum membranes) and functioned properly. Surprisingly, little disruption to endogenous metabolism was observed in the transgenic plants expressing medium levels of dhurrin, and accumulation of pathway intermediates was not observed. The implication is that the plastic nature of plant metabolism can accommodate and regulate activity in new biosynthetic pathways that are introduced. The resulting plants included individuals producing levels of dhurrin similar to sorghum plants in leaf tissue (up to 4 mg/g fresh weight) and produced hydrogen cyanide on tissue damage (due to the hydrolysis of dhurrin by an endogenous *Arabidopsis* enzyme). The dhurrin-expressing plants showed enhanced resistance to attack by the flea beetle *Phyllotreta nemorum*, a specialist feeder on crucifers; adult beetles avoided feeding on dhurrin-expressing leaves when offered a choice, and larvae under no-choice conditions either failed to initiate feeding, or on initiating feeding showed a significant level of mortality. These initial results clearly imply that production of high levels of dhurrin in transgenic *Arabidopsis* caused phenotypic abnormalities, but subsequent refining of the technology allowed accumulation of dhurrin at up to 4% dry weight in *Arabidopsis* tissues without deleterious effects on plant growth [205]; expression levels of the UDP-glycosyl transferase must be high enough to prevent accumulation of the *p*-hydroxymandelonitrile intermediate.

Although these results represent science of the highest quality, this method is of marginal usefulness for crop protection as it stands, due to the dhurrin end product being toxic to higher organisms, due to the production of hydrogen cyanide when it is hydrolyzed. Worse, many insect herbivores, particularly those which have a polyphagous feeding habit, can detoxify cyanide [206]. However, the feasibility of engineering secondary metabolism in crop plants has now been established. Expression of the cassava cyanogenic glycosides, linamarin and lotaustralin (derived from valine and isoleucine respectively), has also been achieved in *Arabidopsis* [207], and grape vine root cultures have been engineered to produce dhurrin [208], although in this case no protection against root aphids was observed. Other types of secondary metabolites have also been exploited; production of the alkaloid caffeine from its precursor xanthosine in tobacco was achieved by the introduction of three genes encoding *N*-methyl transferases [209]. The resulting plants contained up to 5 µg/g fresh weight caffeine in leaves, and showed a strong feeding deterrent effect toward a generalist lepidopteran herbivore, *Spodoptera litura*. An alternative approach to modifying secondary metabolism was taken by [210], who introduced a gene encoding β-glucosidase from *Aspergillus niger* into tobacco plants, and demonstrated that transgenic plants expressing the enzyme had insecticidal activity toward whiteflies (*Bemisia* spp.) and dipterans (flies), putatively due to hydrolysis of unidentified glycosides in the plant (although the greater density of secretory trichomes observed in transgenic plants may also have been significant). Further developments in this area can be expected.

Besides engineering, secondary metabolism to produce defensive compounds normally present in other plant species, the biosynthetic capacity of plants can be used to produce a variety of volatile secondary compounds used for communication. Better understanding of the terpenoid biosynthesis pathways has led to the production of a number of transgenic plants with altered volatile composition (reviewed by Aharoni et al. [211]). Suppression of expression of a cytochrome P450 oxidase gene expressed in trichomes by RNAi led to transgenic tobacco plants which deterred aphid colonization [212], due to the final step in production of the diterpenoid cembratriene-diol being

blocked, resulting in accumulation of the precursor, cembratriene-ol. These compounds are both volatile and components of trichome secretions. Transgenic *Arabidopsis* plants constitutively over-expressing a dual linalool/nerolidol synthase in plastids produced significant amounts of linalool, both as a free alcohol (volatile) and as glycosylated derivatives, and were repellent to aphids (*M. persicae*) when tested in a choice experiment [213]. Modifications to isoprenoid synthesis in *Arabidopsis* have also been shown to attract predatory mites, which could protect plants by destroying pests [214]. This strategy of attracting natural enemies to pests has also been exemplified by transforming *Arabidopsis* with the maize terpene synthase gene TPS10, which is responsible for producing sesquiterpene volatiles emitted by maize. The resulting plants emitted the volatiles normally produced in maize and attracted parasitoid wasps which attack maize pests [215]. A different approach to utilizing terpene production in transgenic plants exploits the activity of the sesquiterpene (E)-β-farnesene as an alarm pheromone in aphids, which causes cessation of feeding and avoidance, as well as acting as an attractant for aphid predators and parasitoids [216]. *Arabidopsis* was transformed with an (E)-β-farnesene synthase gene from mint, under control of a constitutive promoter (CaMV 35S); resulting plants produced (E)-β-farnesene as a volatile. The transgenic plants showed significant levels of aphid deterrence in choice experiments, and were attractive to the aphid parasitoid *Diaeretiella rapae*. Experiments which engineer the volatiles emitted by plants are an exciting area of research at present, which has established the role that volatiles emitted by plants play in the interactions between plants, herbivores, and natural enemies at the tritrophic level. This technology has yet to show that it is a practical method for crop protection in the field, but practical applications look likely to follow.

Some Novel Approaches

Many other approaches to engineering insect resistance in transgenic plants have been proposed, and progressed to varying degrees. The following section gives an overview of some of the most promising of these approaches, which have been taken forward to the stage of demonstrating feasibility by producing insect-resistant plants.

Of necessity, many other interesting ideas have had to be omitted, such as transformation of plants with transcription factors which alter gene expression [217, 218], or the use of transgenic plants expressing potentially toxic proteins from insects [219] or insect peptide hormones [220]. Despite the lack of commercial deployment of any of the insect-resistant transgenic plant other than those expressing proteins derived from *Bt*, this field of research is active and new approaches will continue to be put forward and evaluated.

***Photorhabdus luminescens* Insecticidal Proteins**

Photorhabdus luminescens is an enterobacterial symbiont of entomophagous (insecticidal) nematodes of *Heterorhabditis* species, used for small-scale biological control of insect pests. The bacteria are present in the nematode gut, and when nematodes enter an insect host, bacterial cells are released into the insect circulatory system. The bacterial cells release toxins which cause cell death, leading to a lethal septicemia, providing a substrate for both bacteria and nematodes to grow on [221, 222]. The toxins are present as high-molecular-weight (M_r approx. 10^6) complexes, which are toxic when injected or fed to insects from four major orders of agricultural pests. The complex has been separated into four components, encoded by genetic loci *tca*, *tcb*, *tcc*, and *tcd*; the products of *tca* and *tcd* are toxic individually when fed to lepidopteran larvae. The mechanism of action of the toxins remains unresolved. Subsequent investigation has shown that *Photorhabdus* contains a large number of potentially insecticidal components, some of which are only toxic by injection, whereas others are orally toxic (reviewed by French-Constant [223]); a variety of mechanisms of toxicity, including promotion of apoptosis, seems to be exploited by the bacterium. This presence of a reservoir of redundant insecticidal activities, reminiscent of the situation in *Bacillus thuringiensis*, led to *Photorhabdus* being put forward as a successor to *Bt* as a source of insecticidal genes for expression in transgenic plants.

In order to be able to exploit insecticidal genes, investigators have sought to isolate single toxic proteins from *Photorhabdus*. Two proteins, designated toxin A and toxin B, were isolated from culture supernatant and shown to be orally toxic [224]. They exist as high-molecular-weight complexes (approx. 860 kDa) in

solution, and each consist of two polypeptides, 201 and 63 kDa molecular weight. The mature polypeptides are produced from single precursor protoxin polypeptides of 283 kDa by proteolysis by endogenous bacterial proteinases. The 283 kDa protoxin A is the product of a gene designated *tcdA* in *Photorhabdus*, which has been cloned and assembled into expression constructs for use in transgenic plants. Expression levels of mRNA and protein were improved by adding 5' and 3' UTR sequences from a tobacco osmotin gene, but the coding sequence was not reengineered. Expression in transgenic *Arabidopsis* gave plants that contained intact protoxin, with a range of expression levels [225]; expression of toxin A at levels above 0.07% of total soluble protein in leaves gave almost complete protection against larvae of the lepidopteran tobacco hornworm (*M. sexta*). The toxin is not species specific, and leaf extracts were also toxic to the coleopteran corn rootworm (*Diabrotica undecimpunctata*). Commercial development of this technique is highly likely.

Entomophagous nematodes of *Steinernema* species also contain mutualistic bacteria, of *Xenorhabdus* species, which produce insecticidal toxins. These proteins could also be exploited to produce insect resistance in transgenic plants, but have not yet received as much attention as *Photorhabdus* toxins [223].

Cholesterol Oxidase

The identification of a protein from *Streptomyces* that was highly insecticidal to larvae of the coleopteran pest cotton boll weevil (*Anthonomus grandis*) resulted from a screening program assaying culture filtrates of different bacterial species [226]. The protein, which was toxic at levels comparable to a *Bt* three-domain Cry protein, was identified as a cholesterol oxidase. It was able to lyse the midgut epithelium in the insect. The mechanism of action involves the activity of the enzyme, since no activity is seen in lepidopteran larvae where the gut pH is high, and the enzyme has low activity, but may also involve effects on membrane-bound alkaline phosphatase [227]. Oxidation of membrane sterols such as cholesterol in the insect gut epithelium can destabilize membranes, leading to cell lysis as observed. However, expression of this protein in transgenic plants could prove problematic, since

it is equally capable of oxidizing sterols in plant cell membranes. The encoding gene for the cholesterol oxidase was isolated, and assembled into expression constructs containing either the complete coding sequence, the mature protein coding sequence, or the coding sequence fused to a chloroplast targeting peptide from the *Arabidopsis ribulose* biphosphate carboxylase (RuBisCO) small subunit gene [228]. No codon optimization was carried out. Transgenic tobacco plants were produced by transformation of the nuclear genome, and all constructs were shown to result in synthesis and accumulation of active enzyme. The constructs which omitted the chloroplast targeting peptide caused protein to accumulate in the cytoplasm, and these plants were developmentally abnormal, possibly as a result of interference with plant sterol hormone signaling pathways. Plants in which the enzyme was localized in chloroplasts were phenotypically normal. Leaf tissue from all transgenic plants was toxic to boll weevil larvae when fed as a component of an artificial diet.

This work does not seem to have been progressed beyond the stage of a demonstration of concept, and no further references to it are present in the scientific literature. This gene would seem a good candidate for introduction into the chloroplast genome to engineer insect resistance, although potential effects on chloroplast membrane systems would remain a drawback.

Avidin as an Insecticidal Protein

Exploitation of the biotin-binding properties of the avian egg white protein avidin (and its bacterial functional homologue, streptavidin) in a variety of biochemical techniques has obscured its role as a defensive protein, which is toxic to bacteria. The antibacterial activity is based on its essentially irreversible binding of biotin, leading to this essential enzyme cofactor being unavailable. The insecticidal activity of avidin was recognized in 1993, when assays carried out in artificial diet showed toxicity to coleopteran and lepidopteran larvae at levels as low as 10 ppm in diet (estimated as of the order of 0.01% of total protein), although the level necessary to show toxicity was up to 100x higher for other pest species. The toxic effect was eliminated by addition of biotin to diets, suggesting

that the mechanism of avidin insecticidal activity is also through biotin sequestration. Both growth reduction and mortality were observed, and the suggestion was made that gene constructs expressing avidin could provide protection against insect pests in transgenic plants [229]. Subsequent assays confirmed that susceptibility to avidin as an insecticide varies widely between different insect species, and that biotin carried over in the egg between generations had a significant effect on subsequent avidin toxicity [230].

Initial reports of expression of avidin in transgenic maize were focused on producing the protein as a high-value product [231]. An expression construct containing a codon-optimized avidin coding sequence with an N-terminally fused signal peptide from barley α -amylase, driven by the maize ubiquitin-1 promoter, resulted in expression levels of avidin of >2.0% of total protein in seed. Seed from these plants was subsequently bioassayed for resistance to larvae of three different coleopteran storage pests, including red flour beetle (*Tribolium castaneum*), with 100% mortality at avidin levels above 100 ppm of seed (approx. 0.1% of total protein). However, not all pests were as susceptible; larvae of the larger grain borer, *Prostephanus truncatus*, were effectively insensitive to avidin, whether added to artificial diet or expressed in transgenic plant material. The engineered maize was nontoxic to mice over 21 days [232]. Subsequent reports confirmed the insecticidal effects of avidin expressed in transgenic plants: these include protection of tobacco against noctuid lepidopterans [233], using vacuolar targeting sequences from potato proteinase inhibitors to direct avidin accumulation in the vacuole at levels up to 1.5% of total leaf protein [234]; protection of apple against lepidopteran pests [235]; and protection of rice against coleopteran stored grain pests, using a similar approach to that used for maize [236]. Targeting of the foreign protein to vacuolar or similar compartments is important; expression of streptavidin in tomato using plant and bacterial signal peptides and strong promoters led to developmental abnormalities in the plants, which could be corrected by topical application of biotin, suggesting that sequestration of cellular biotin is equally detrimental for plants as well as insects [237].

Despite many promising results, this technology appears to have failed to gain any acceptance for agricultural crops, as illustrated by a recent study in which seed meal from transgenic avidin-expressing maize was tested as an insecticide for topical application to stored maize [238]. Studies have shown that avidin can increase the protection afforded by *Bt* expression in transgenic plants against insect pests which have limited susceptibility to the toxin (e.g., potato expressing Cry3A; [239]), but it is clear that little further development in this area is taking place.

RNA Interference Using Double-Stranded RNA

Downregulation of gene expression by double-stranded RNA (dsRNA) corresponding to part or all of a specific gene transcript has been used as a research technique in insect genetics since 1998. The method has been based on delivery of synthetic dsRNA produced in vitro by injection into insect cells or tissues, which is clearly not practical for applications in crop protection. However, recent results have shown that dsRNA can be introduced into insects as a component of artificial diet, and is effective in downregulating genes normally expressed in gut tissue. This technique has been used to downregulate the production of a gut carboxylesterase in larvae of the lepidopteran *Epiphyas postvittana* (light brown apple moth; [240]), leading to suppression of mRNA in the insect. More significantly, two recent papers show that dsRNA can be delivered to insect pests by expression in plant material, and that this can lead to an insecticidal effect when pests are exposed to plants. Transgenic tobacco and *Arabidopsis* plant material expressing dsRNA directed against a cotton bollworm detoxification enzyme (cytochrome P450 gene CYP6AE14) for gossypol suppressed expression of the gene, and caused the insect to become more sensitive to gossypol in the diet, leading to reduced performance compared to controls [241]. A similar technique was used to suppress expression of a V-type ATPase in larvae of the coleopteran *Diabrotica virgifera virgifera* (Western corn rootworm); transgenic corn plants producing dsRNA directed against this gene showed protection against feeding damage by the insect [242]. The feasibility of using dsRNA in crop protection strategies has thus been demonstrated. This approach holds great promise for future development,

as it allows a wide range of potential targets for suppression of gene expression in the insect to be exploited.

Insect-Resistant Genetically Engineered Crops and Sustainability

The success of Bt-expressing crops in the field has been a direct result of taking “sustainability” into account in their introduction, particularly with respect to managing the emergence of pest resistance to the toxins through the refuge strategy, as described earlier. Even organizations hostile to Genetic Engineering technology, such as organic growers in the USA, have reported that *Bt* cotton and corn have reduced insecticide usage significantly (by up to 0.2 kg/ha/year), showing that these crops are compatible with the goals of “sustainable” agriculture [243].

The “sustainability” of transgenic insect-resistant crops has also been examined in terms of potential effects on the wider ecosystem in which the plants are grown. Numerous studies have been carried out to effects on predators and parasites at the third trophic level, and on nontarget insects and other invertebrates. Some initial reports which did report negative effects were based on dubious assumptions, or used experimental designs which had little relevance to field conditions (e.g., the supposed threat to monarch butterflies posed by transgenic *Bt* corn; reviewed by Gatehouse et al. [244]). Nevertheless, it must be the case that if a pest population is decreased as a result of endogenous resistance in crops, then there will be a “knock on” effect to the wider ecosystem, and particularly to predators and parasites of the pest species, when the resistant crop is compared to a nonresistant one that is not treated with pesticide. However, this is not a realistic comparison, since in agricultural practice a crop that does not have endogenous resistance is treated with exogenous insecticides. The use of the refuge strategy allows significant pest populations to be present, and thus can support both beneficial insects which attack the pest, and a wider ecosystem, which would be destroyed by exogenous insecticide application.

Looking to the future, wider use of insect-resistant transgenic crops could contribute positively to “sustainability” in agriculture in general, by further decreasing

insecticide usage and thereby decreasing energy inputs. However, the “sustainability” of the insect-resistant crops themselves is going to come under increasing pressure, as less controlled deployment of insect-resistant plants evades the present compulsory use of the refuge strategy, and use of crop varieties with multiple *Bt* toxins renders the refuge strategy apparently less necessary to prevent pest resistance to *Bt* toxins developing. Field resistance to *Bt* crops has been observed recently (reviewed by Tabashnik et al. [245]), but is manageable using existing practices, or modifications of them. The sustainability of relying on one mechanism of crop protection can be questioned, especially as plants in general have evolved mixed defense strategies [246]. In the longer term, a wider range of strategies for producing insect-resistant plants is going to be necessary, not only to deal with the potential for nonspecific resistance to *Bt* toxins, but to extend the range of crop pests that can be targeted, and further reduce the application of pesticides.

Future Directions

After 20 years, insect-resistant transgenic crops have been a greater success in some ways than the early experiments suggested, but have failed to meet all the hopes that were initially raised. The success is self-evident when the widescale adoption of the technology in certain crops such as cotton and maize is considered, and documented evidence of reductions in damage to human health and the environment as a result of decreases in the use of exogenously applied pesticides. The failure does not lie in any technical shortcomings in the science, although improvements and new strategies are always possible; it lies in a failure to disseminate the technology as widely as should have been the case, so that it remains largely in the hands of commercial organizations, and is limited to the major crops. Is it an unrealistic hope to anticipate that after another 20 years, amateur gardeners in developed countries will be able to choose to buy seed to grow genetically engineered cabbages, which will be resistant to cabbage white butterfly larvae, in their allotments and gardens? Or that rural farmers in developing countries will have free access to engineered rice varieties, suitable for their growth conditions, that are resistant to pests such as stemborers? Both these aims have been scientifically

achievable for at least the last 10 years, and it is surely about time that a more rational approach, which cuts through both the largely futile debate about the rights and wrongs of plant genetic engineering, and the protectionism of agrochemical companies, was taken to address the looming problem of producing enough crops to meet humanity's needs.

Bibliography

Primary Literature

1. Barton KA, Whitely HR, Yang N-S (1987) *Bacillus thuringiensis* δ -endotoxin expressed in transgenic *Nicotiana tabacum* provides resistance to Lepidopteran insects. *Plant Physiol* 85: 1103–1109
2. Fischhoff DA, Bowdish KS, Perlak FJ, Marrone PG, McCormick SH, Niedermeyer JG, Dean DA, Kusano-Kretzmer K, Mayer EJ, Rochester DE, Rogers SG, Fraley RT (1987) Insect tolerant transgenic tomato plants. *Bio/Technology* 5:807–813
3. Vaeck M, Reynaerts A, Hofte H, Jansens S, De Beuckeleer M, Dean C, Zabeau M, Van Montagu M, Leemans J (1987) Transgenic plants protected from insect attack. *Nature (London)* 328:33–37
4. Toenniessen GH, O'Toole JC, DeVries J (2003) Advances in plant biotechnology and its adoption in developing countries. *Curr Opin Plant Biol* 6:191–198
5. Shelton AM, Zhao J-Z, Roush RT (2002) Economic, ecological, food safety, and social consequences of the deployment of *Bt* transgenic plants. *Annu Rev Entomol* 47:845–881
6. Aronson AI, Shai Y (2001) Why *Bacillus thuringiensis* insecticidal toxins are so effective: unique features of their mode of action. *FEMS Microbiol Lett* 195:1–8
7. de Maagd RA, Bravo A, Crickmore N (2001) How *Bacillus thuringiensis* has evolved specific toxins to colonize the insect world. *Trends Genet* 17:193–199
8. Damgaard PH, Hansen BM, Pedersen JC, Eilenberg J (1997) Natural occurrence of *Bacillus thuringiensis* on cabbage foliage and in insects associated with cabbage crops. *J Appl Microbiol* 82:253–258
9. Bizzarri MF, Bishop AH (2007) Recovery of *Bacillus thuringiensis* in vegetative form from the phylloplane of clover (*Trifolium hybridum*) during a growing season. *J Inverteb Pathol* 94: 38–47
10. Bernhard K, Jarrett P, Meadows M, Butt J, Ellis DJ, Roberts GM, Pauli S, Rodgers P, Burges HD (1997) Natural isolates of *Bacillus thuringiensis*: worldwide distribution, characterization, and activity against insect pests. *J Inverteb Pathol* 70:59–68
11. Crickmore N, Zeigler DR, Feitelson J, Schnepf E, Van Rie J, Lereclus D, Baum J, Dean DH (1998) Revision of the nomenclature for the *Bacillus thuringiensis* pesticidal crystal proteins. *Microbiol Mol Biol Rev* 62:807–813
12. Berry C, O'Neil S, Ben-Dov E, Jones AF, Murphy L, Quail MA, Holden MTG, Harris D, Zaritsky A, Parkhill J (2002) Complete

- sequence and organization of pBtoxis, the toxin-coding plasmid of *Bacillus thuringiensis* subsp. *israeliensis*. Appl Environ Microbiol 68:5082–5095
13. Parker MW, Feil SC (2005) Pore-forming protein toxins: from structure to function. Prog Biophys Mol Biol 88:91–142
 14. Li J, Carrol J, Ellar DJ (1991) Crystal structure of insecticidal δ -endotoxin from *Bacillus thuringiensis* at 2.5 Å resolution. Nature 353:815–821
 15. Grochulski P, Masson L, Borisova S, Pusztai-Carey M, Schwartz JL, Brousseau R, Cygler M (1995) *Bacillus thuringiensis* CryIA(a) insecticidal toxin: crystal structure and channel formation. J Mol Biol 254:447–464
 16. Morse RJ, Yamamoto T, Stroud RM (2001) Structure of Cry2Aa suggests an unexpected receptor binding epitope. Structure 9:409–417
 17. Galitsky N, Cody V, Wojtczak A, Ghosh D, Luft JR, Pangborn W, English L (2001) Structure of the insecticidal bacterial δ -endotoxin CryBb1 of *Bacillus thuringiensis*. Acta Crystallogr D 57:1101–1109
 18. Boonserm P, Mo M, Angsuthanasombat C, Lescar J (2006) Structure of the functional form of the mosquito larvicidal Cry4Aa toxin from *Bacillus thuringiensis* at a 2.8-Ångstrom resolution. J Bacteriol 188:3391–3401
 19. Boonserm P, Davis P, Ellar DJ, Li J (2005) Crystal structure of the mosquito-larvicidal toxin Cry4Ba and its biological implications. J Mol Biol 348:363–382
 20. de Maagd RA, Bravo A, Berry C, Crickmore N, Schnepf HE (2003) Structure, diversity and evolution of protein toxins from spore-forming entomopathogenic bacteria. Annu Rev Genet 37:409–433
 21. Bravo A, Sánchez J, Kouskoura T, Crickmore N (2002) N-terminal activation is an essential early step in the mechanism of action of the *B. thuringiensis* Cry1Ac insecticidal toxin. J Biol Chem 277:23985–23987
 22. Bravo A, Gill SS, Soberón M (2007) Mode of action of *Bacillus thuringiensis* Cry and Cyt toxins and their potential for insect control. Toxicon 49:423–435
 23. Knight P, Crickmore N, Ellar DJ (1994) The receptor for *Bacillus thuringiensis* CryIA(c) delta-endotoxin in the brush border membrane of the lepidopteran *Manduca sexta* is aminopeptidase N. Mol Microbiol 11:429–436
 24. Vadlamudi RK, Weber E, Ji I, Ji TH, Bulla LA Jr (1995) Cloning and expression of a receptor for an insecticidal toxin of *Bacillus thuringiensis*. J Biol Chem 270:5490–5494
 25. Valaitis AP, Jenkins JL, Lee MK, Dean DH, Garner KJ (2001) Isolation and partial characterization of Gypsy moth BTR-270, an anionic brush border membrane glycoconjugate that binds *Bacillus thuringiensis* Cry1A toxins with high affinity. Arch Insect Biochem Physiol 46:186–200
 26. Jurat-Fuentes JL, Adang MJ (2004) Characterization of a Cry1Ac-receptor alkaline phosphatase in susceptible and resistant *Heliothis virescens* larvae. Eur J Biochem 271: 3127–3135
 27. Fernández LE, Aimanova KG, Gill SS, Bravo A, Soberón M (2006) A GPI-anchored alkaline phosphatase is a functional midgut receptor of Cry11Aa toxin in *Aedes aegypti* larvae. Biochem J 394:77–84
 28. Krishnamoorthy M, Jurat-Fuentes JL, McNall RJ, Andacht T, Adang MJ (2007) Identification of novel CryIAc binding proteins in midgut membranes from *Heliothis virescens* using proteomic analyses. Insect Biochem Mol Biol 37:189–201
 29. Gahan LJ, Gould F, Heckel DG (2001) Identification of a gene associated with *Bt* resistance in *Heliothis virescens*. Science 293:857–860
 30. Yang YJ, Chen HY, Wu SW, Yang YH, Xu XJ, Wu YD (2006) Identification and molecular detection of a deletion mutation responsible for a truncated cadherin of *Helicoverpa armigera*. Insect Biochem Mol Biol 36:735–740
 31. Xie R, Zhuang M, Ross LS, Gómez I, Oltean DI, Bravo A, Soberón M, Gill SS (2005) Single amino acid mutations in the cadherin receptor from *Heliothis virescens* affect its toxin binding ability to Cry1A toxins. J Biol Chem 280:8416–8425
 32. Tsuda Y, Nakatani F, Hashimoto K, Ikawa S, Matsuura C, Fukada T, Sugimoto K, Himeno M (2003) Cytotoxic activity of *Bacillus thuringiensis* Cry proteins on mammalian cells transfected with cadherin-like Cry receptor gene of *Bombyx mori* (silkworm). Biochem J 369:697–703
 33. Soberón M, Pardo-López L, López I, Gómez I, Tabashnik BE, Bravo A (2007) Engineering modified *Bt* toxins to counter insect resistance. Science 318:1640–1642
 34. Herrero S, Gechev T, Bakker PL, Moar WJ, de Maagd RA (2005) *Bacillus thuringiensis* Cry1Ca-resistant *Spodoptera exigua* lacks expression of one of four aminopeptidase N genes. BMC Genomics 6:96
 35. Rajagopal R, Sivakumar S, Agrawal N, Malhotra P, Bhatnagar RK (2002) Silencing of midgut aminopeptidase N of *Spodoptera litura* by double-stranded RNA establishes its role as *Bacillus thuringiensis* toxin receptor. J Biol Chem 277: 46849–46851
 36. Sivakumar S, Rajagopal R, Venkatesh GR, Srivastava A, Bhatnagar RK (2007) Knockdown of aminopeptidase-N from *Helicoverpa armigera* larvae and in transfected Sf21 cells by RNA interference reveals its functional interaction with *Bacillus thuringiensis* insecticidal protein Cry1Ac. J Biol Chem 282:7312–7319
 37. Gill M, Ellar D (2002) Transgenic *Drosophila* reveals a functional *in vivo* receptor for the *Bacillus thuringiensis* toxin Cry1Ac1. Insect Mol Biol 11:619–625
 38. Burton SL, Ellar DJ, Li J, Derbyshire DJ (1999) N-acetylgalactosamine on the putative insect receptor aminopeptidase N is recognised by a site on the domain III lectin-like fold of a *Bacillus thuringiensis* insecticidal toxin. J Mol Biol 287:1011–1022
 39. Knight PJK, Carroll J, Ellar DJ (2004) Analysis of glycan structures on the 120 kDa aminopeptidase N of *Manduca sexta* and their interactions with *Bacillus thuringiensis* CryIAc toxin. Insect Biochem Mol Biol 34:101–112
 40. de Maagd RA, Bakker PL, Masson L, Adang MJ, Sangadala S, Stiekema W, Bosch D (1999) Domain III of the *Bacillus thuringiensis* delta-endotoxin Cry1Ac is involved in binding

- to *Manduca sexta* brush border membranes and to its purified aminopeptidase N. *Mol Microbiol* 31:463–471
41. Jenkins JL, Lee MK, Valaitis AP, Curtiss A, Dean DH (2000) Bivalent sequential binding model of a *Bacillus thuringiensis* toxin to gypsy moth aminopeptidase N receptor. *J Biol Chem* 275: 14423–14431
 42. Jimenez-Juarez N, Munoz-Garay C, Gomez I, Saab-Rincon G, Damian-Almazo JY, Gill SS, Soberon M, Bravo A (2007) *Bacillus thuringiensis* Cry1Ab mutants affecting oligomer formation are non-toxic to *Manduca sexta* larvae. *J Biol Chem* 282:21222–21229
 43. Rausell C, García-Robles I, Sánchez J, Muñoz-Garay C, Martínez-Ramírez AC, Real MD, Bravo A (2004) Role of toxin activation on binding and pore formation activity of the *Bacillus thuringiensis* Cry3 toxins in membranes of *Leptinotarsa decemlineata* [Say]. *Biochem Biophys Acta* 1660:99–105
 44. Gómez I, Sánchez J, Miranda R, Bravo A, Soberón M (2002) Cadherin-like receptor binding facilitates proteolytic cleavage of helix a-1 in domain I and oligomer pre-pore formation of *Bacillus thuringiensis* Cry1Ab toxin. *FEBS Lett* 513:242–246
 45. Chen J, Hua G, Jurat-Fuentes JL, Abdullah MA, Adang MJ (2007) Synergism of *Bacillus thuringiensis* toxins by a fragment of a toxin-binding cadherin. *Proc Natl Acad Sci USA* 104: 13901–13906
 46. Parenti P, Morandi P, McGivan JD, Consonnic P, Leonardi G, Giordana B (1997) Properties of the aminopeptidase N from the silkworm midgut (*Bombyx mori*). *Insect Biochem Mol Biol* 27:397–403
 47. Zhuang M, Oltean DI, Gómez I, Pullikuth AK, Soberón M, Bravo A, Gill SS (2002) *Heliothis virescens* and *Manduca sexta* lipid rafts are involved in Cry1A toxin binding to the midgut epithelium and subsequent pore formation. *J Biol Chem* 277:13863–13872
 48. Bravo A, Gómez I, Conde J, Muñoz-Garay C, Sánchez J, Zhuang M, Gill SS, Soberón M (2004) Oligomerization triggers differential binding of a pore-forming toxin to a different receptor leading to efficient interaction with membrane microdomains. *Biochem Biophys Acta* 1667:38–46
 49. Pigott CR, Ellar DJ (2007) Role of receptors in *Bacillus thuringiensis* crystal toxin activity. *Microbiol Mol Biol Rev* 71:255–281
 50. Sacchi VF, Wolfsberger MG (1996) Amino acid absorption. In: Lehane MJ, Billingsley PF (eds) *Biology of the insect midgut*. Chapman and Hall, London, pp 265–292
 51. Zhang X, Candas M, Griko NB, Taussig R, Bulla LA Jr (2006) A mechanism of cell death involving an adenylyl cyclase/PKA signaling pathway is induced by the Cry1Ab toxin of *Bacillus thuringiensis*. *Proc Natl Acad Sci USA* 103:9897–9902
 52. Moellenbeck DJ, Peters ML, Bing JW, Rouse JR, Higgins LS, Sims L, Nevshemal T, Marshall L, Ellis RT, Bystrak PG, Lang BA, Stewart JL, Kouba K, Sondag V, Gustafson V, Nour K, Xu DP, Swenson J, Zhang J, Czaplá T, Schwab G, Jayne S, Stockhoff BA, Narva K, Schnepf HE, Stelman SJ, Poutre C, Koziel M, Duck N (2001) Insecticidal proteins from *Bacillus thuringiensis* protect corn from corn rootworms. *Nat Biotechnol* 19:668–672
 53. Ellis RT, Stockhoff BA, Stamp L, Schnepf HE, Schwab GE, Knuth M, Russell J, Cardineau GA, Narva KE (2002) Novel *Bacillus thuringiensis* binary insecticidal crystal proteins active on western corn rootworm, *Diabrotica virgifera virgifera* LeConte. *Appl Environ Microbiol* 68:1137–1145
 54. Darboux I, Nielsen-LeRoux C, Charles JF, Pauron D (2001) The receptor of *Bacillus sphaericus* binary toxin in *Culex pipiens* (Diptera: Culicidae) midgut: molecular cloning and expression. *Insect Biochem Mol Biol* 31:981–990
 55. Charles JF, Nielsen-LeRoux C, Delecluse A (1996) *Bacillus sphaericus* toxins: molecular biology and mode of action. *Annu Rev Entomol* 41:451–472
 56. Warren GW (1997) Vegetative insecticidal proteins: novel proteins for control of corn pests. In: Carozzi NB, Koziel MG (eds) *Advances in insect control: the role of transgenic plants*. Taylor & Francis, London, UK, pp 109–121
 57. Barth H, Aktories K, Popoff MR, Stiles BG (2004) Binary bacterial toxins: Biochemistry, biology, and applications of common *Clostridium* and *Bacillus* proteins. *Microbiol Mol Biol Rev* 68:373–402
 58. Leuber M, Orlik F, Schiffler B, Sickmann A, Benz R (2006) Vegetative insecticidal protein (Vip1Ac) of *Bacillus thuringiensis* HD201: Evidence for oligomer and channel formation. *Biochemistry* 45:283–288
 59. Estruch JJ, Warren GW, Mullins MA, Nye GJ, Craig JA, Koziel MG (1996) Vip3A, a novel *Bacillus thuringiensis* vegetative insecticidal protein with a wide spectrum of activities against lepidopteran insects. *Proc Natl Acad Sci USA* 93:5389–5394
 60. Yu CG, Mullins MA, Warren GW, Koziel MG, Estruch JJ (1997) The *Bacillus thuringiensis* vegetative insecticidal protein Vip3A lyses midgut epithelium cells of susceptible insects. *Appl Environ Microbiol* 63:532–536
 61. Lee MK, Miles P, Chen JS (2006) Brush border membrane binding properties of *Bacillus thuringiensis* Vip3A toxin to *Heliothis virescens* and *Helicoverpa zea* midguts. *Biochem Biophys Res Commun* 339:1043–1047
 62. Lee MK, Walters FS, Hart H, Palekar N, Chen JS (2003) Mode of action of the *Bacillus thuringiensis* vegetative insecticidal protein Vip3A differs from that of Cry1Ab delta-endotoxin. *Appl Environ Microbiol* 69:4648–4657
 63. Fang J, Xu X, Wang P, Zhao J-Z, Shelton AM, Cheng J, Feng M-G, Shen Z (2007) Characterization of chimeric *Bacillus thuringiensis* Vip3 toxins. *Appl Environ Microbiol* 73:956–961
 64. Li J, Pandelakis AK, Ellar DJ (1996) Structure of the mosquitocidal δ -endotoxin CytB from *Bacillus thuringiensis* sp. *kyushuensis* and implications for membrane pore formation. *J Mol Biol* 257:129–152
 65. Du J, Knowles BH, Li J, Ellar DJ (1999) Biochemical characterization of *Bacillus thuringiensis* cytolytic toxins in association with a phospholipid bilayer. *Biochem J* 338:185–193
 66. Promdonkoy B, Ellar DJ (2003) Investigation of the pore-forming mechanism of a cytolytic δ -endotoxin from *Bacillus thuringiensis*. *Biochem J* 374:255–259

67. Koni PA, Ellar DJ (1994) Biochemical characterization of *Bacillus thuringiensis* cytolytic δ -endotoxins. *Microbiology* 140:1869–1880
68. Wirth MC, Georgiopoulos GP, Federici BA (1997) CytA enables CryIV endotoxins of *Bacillus thuringiensis* to overcome high levels of CryIV resistance in the mosquito, *Culex quinquefasciatus*. *Proc Natl Acad Sci USA* 94:10536–10540
69. Pérez C, Fernández LE, Sun J, Folch JL, Gill SS, Soberón M, Bravo A (2005) *Bacillus thuringiensis* subsp. *israelensis* Cyt1Aa synergizes Cry11Aa toxin by functioning as a membrane-bound receptor. *Proc Natl Acad Sci USA* 102:18303–18308
70. Mazier M, Pannetier C, Tourneur J, Jouanin L, Giband M (1997) The expression of *Bacillus thuringiensis* toxin genes in plant cells. *Biotechnol Annu Rev* 3:313–347
71. Zheng SJ, Henken B, de Maagd RA, Purwito A, Krens FA, Kik C (2005) Two different *Bacillus thuringiensis* toxin genes confer resistance to beet armyworm (*Spodoptera exigua* Hubner) in transgenic *Bt*-shallots (*Allium cepa* L.). *Transgenic Res* 14: 261–272
72. Fujimoto H, Itoh K, Yamamoto M, Kyoizuka J, Shimamoto K (1993) Insect resistant rice generated by introduction of a modified δ -endotoxin gene of *Bacillus thuringiensis*. *Bio/Technology* 11:194–200
73. Wunn J, Kloti A, Burkhardt PK, Biswas GCG, Launis K, Iglesias VA, Potrykus I (1996) Transgenic Indica rice breeding line IR58 expressing a synthetic cryIA(b) gene from *Bacillus thuringiensis* provides effective insect pest control. *Bio/Technology* 14:171–176
74. Nayak P, Basu D, Das S, Basu A, Ghosh D, Ramakrishnan NA, Ghosh M, Sen SK (1997) Transgenic elite indica rice plants expressing CryIAC delta-endotoxin of *Bacillus thuringiensis* are resistant against yellow stem borer (*Scirpophaga incertulas*). *Proc Natl Acad Sci USA* 94:2111–2116
75. Vaughn T, Cavato T, Brar G, Coombe T, DeGooyer T, Ford S, Groth M, Howe A, Johnson S, Kolacz K, Pilcher C, Purcell J, Romano C, English L, Pershing J (2005) A method of controlling corn rootworm feeding using a *Bacillus thuringiensis* protein expressed in transgenic maize. *Crop Sci* 45:931–938
76. Breiter JC, Cordero MJ, Royer M, Meynard D, San Segundo B, Guiderdoni E (2001) The-689/+197 region of the maize protease inhibitor gene directs high level, wound-inducible expression of the *cry1B* gene which protects transgenic rice plants from stemborer attack. *Mol Breed* 7:259–274
77. Breiter JC, Vassal JM, Catala MD, Meynard D, Marfa V, Mele E, Royer M, Murillo I, San Segundo B, Guiderdoni E, Messegueur J (2004) *Bt* rice harbouring *Cry* genes controlled by a constitutive or wound-inducible promoter: protection and transgene expression under Mediterranean field conditions. *Plant Biotechnol J* 2:417–430
78. Miklos JA, Alibhai MF, Bledig SA, Connor-Ward DC, Gao A-G, Holmes BA, Kolacz KH, Kabuye VT, MacRae TC, Paradise MS, Toedebusch AS, Harrison LA (2007) Characterization of soybean exhibiting high expression of a synthetic *Bacillus thuringiensis* cry1A transgene that confers a high degree of resistance to lepidopteran pests. *Crop Sci* 47:148–157
79. Perlak FJ, Fuchs RL, Dean DA, McPherson SL, Fischhoff DA (1991) Modification of the coding sequence enhances plant expression of insect controlling protein genes. *Proc Natl Acad Sci USA* 88:3324–3328
80. De Rocher EJ, Vargo-Gogola TC, Diehn SH, Green PJ (1998) Direct evidence for rapid degradation of *Bacillus thuringiensis* toxin mRNA as a cause of poor expression in plants. *Plant Physiol* 117:1445–1461
81. Murray EE, Rocheleau T, Eberle M, Stock C, Sekar V, Adang M (1991) Analysis of unstable RNA transcripts of insecticidal crystal protein genes of *Bacillus thuringiensis* in transgenic plants and electroporated protoplasts. *Plant Mol Biol* 16: 1035–1050
82. Diehn PJ, Chiu SH, De Rocher WL, Green EJ (1998) Premature polyadenylation at multiple sites within a *Bacillus thuringiensis* toxin gene-coding region. *Plant Physiol* 117:1433–1443
83. Misztal LH, Mostowska A, Skibinska M, Bajsa J, Musial WG, Jarmolowski A (2004) Expression of modified Cry1Ac gene of *Bacillus thuringiensis* in transgenic tobacco plants. *Mol Biotechnol* 26:17–26
84. McBride KE, Svab Z, Schaefer DJ, Hogan PS, Stalker KM, Maliga P (1995) Amplification of a chimeric *Bacillus* gene in chloroplasts leads to an extraordinary level of an insecticidal protein in tobacco. *Bio/Technology* 13:362–365
85. Bock R (2001) Transgenic plastids in basic research and plant biotechnology. *J Mol Biol* 312:425–438
86. Maliga P (2003) Progress towards commercialization of plastid transformation technology. *Trends Biotechnol* 21:20–28
87. Daniell H, Khan MS, Allison L (2002) Milestones in chloroplast genetic engineering: an environmentally friendly era in biotechnology. *Trends Plant Sci* 7:84–91
88. Chakrabarti SK, Lutz KA, Lertwiriawong B, Svab Z, Maliga P (2006) Expression of the *cry9Aa2 B.t.* gene in tobacco chloroplasts confers resistance to potato tuber moth. *Transgenic Res* 15:481–488
89. De Cosa B, Moar W, Lee SB, Miller M, Daniell H (2001) Overexpression of the *Bt* cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. *Nat Biotechnol* 19:71–74
90. Kota M, Daniell H, Varma S, Garczynski SF, Gould F, Moar WJ (1999) Overexpression of the *Bacillus thuringiensis* (*Bt*) Cry2Aa2 protein in chloroplasts confers resistance to plants against susceptible and *Bt*-resistant insects. *Proc Natl Acad Sci USA* 96:1840–1845
91. Reddy VS, Leelavathi S, Selvapandian A, Raman R, Giovanni F, Shukla V, Bhatnagar RK (2002) Analysis of chloroplast transformed tobacco plants with *cry1Ia5* under rice *psbA* transcriptional elements reveal high level expression of *Bt* toxin without imposing yield penalty and stable inheritance of transplastome. *Mol Breed* 9:259–269
92. Dufourmantel N, Tissot G, Goutorbe F, Garçon F, Muhr C, Jansens S, Pelissier B, Peltier G, Dubald M (2005) Generation and analysis of soybean plastid transformants expressing *Bacillus thuringiensis* Cry1Ab protoxin. *Plant Mol Biol* 58:659–668

93. Chrispeels MJ, Sadava DE (1994) Plants, genes and agriculture. Jones and Bartlett, London
94. Gould F (1998) Sustainability of transgenic insecticidal cultivars: integrating pest genetics and ecology. *Annu Rev Entomol* 43:701–726
95. Gould F, Anderson A, Jones A, Sumerford D, Heckel DG, Lopez J, Micinski S, Leonard R, Laster M (1997) Initial frequency of alleles for resistance to *Bacillus thuringiensis* toxins in field populations of *Heliothis virescens*. *Proc Natl Acad Sci USA* 94:3519–3523
96. Andreadis SS, Álvarez-Alfagene A, Sánchez-Ramos I, Stodola TJ, Andow DA, Milonas PG, Savopoulou-Soultani M, Castániera P (2007) Frequency of resistance to *Bacillus thuringiensis* toxin Cry1Ab in Greek and Spanish population of *Sesamia nonagrioides* (Lepidoptera: Noctuidae). *J Econ Entomol* 100:195–201
97. Tabashnik BE, Patin AL, Dennehy TJ, Liu Y-B, Carrière Y, Sims M, Antilla L (2000) Frequency of resistance to *Bacillus thuringiensis* in field populations of pink bollworm. *Proc Natl Acad Sci USA* 97:12980–12984
98. Tabashnik BE, Dennehy TJ, Carrière Y (2005) Delayed resistance to transgenic cotton in pink bollworm. *Proc Natl Acad Sci USA* 102:15389–15393
99. Cristofolletti PT, de Sousa FAM, Rahbe Y, Terra WR (2006) Characterization of a membrane-bound aminopeptidase purified from *Acyrtosiphon pisum* midgut cells. *FEBS J* 273:5574–5588
100. Stewart SD, Adamczyk JJ, Knighten KS, Davis FM (2001) Impact of *Bt* cottons expressing one or two insecticidal proteins of *Bacillus thuringiensis* Berliner on growth and survival of noctuid (Lepidoptera) larvae. *J Econ Entomol* 94:752–760
101. Chitkowski RL, Turnipseed SG, Sullivan MJ, Bridges WC (2003) Field and laboratory evaluations of transgenic cottons expressing one or two *Bacillus thuringiensis* var. *kurstaki* Berliner proteins for management of noctuid (Lepidoptera) pests. *J Econ Entomol* 96:755–762
102. Zhao JZ, Cao J, Li YX, Collins HL, Roush RT, Earle ED, Shelton AM (2003) Transgenic plants expressing two *Bacillus thuringiensis* toxins delay insect resistance evolution. *Nat Biotechnol* 21:1493–1497
103. Christou P, Capell T, Kohli A, Gatehouse JA, Gatehouse AMR (2006) Recent developments and future prospects in insect pest control in transgenic crops. *Trends Plant Sci* 11:302–308
104. Gahan LJ, Ma YT, Coble MLM, Gould F, Moar WJ, Heckel DG (2005) Genetic basis of resistance to Cry1Ac and Cry2Aa in *Heliothis virescens* (Lepidoptera: Noctuidae). *J Econ Entomol* 98:1357–1368
105. Bashir K, Husnain T, Fatima T, Riaz N, Makhdoom R, Riazuddin S (2005) Novel indica basmati line (B-370) expressing two unrelated genes of *Bacillus thuringiensis* is highly resistant to two lepidopteran insects in the field. *Crop Prot* 24:870–879
106. Dively GP (2005) Impact of transgenic VIP3A x Cry1Ab lepidopteran-resistant field corn on the nontarget arthropod community. *Environ Entomol* 34:1267–1291
107. Han LZ, Wu KM, Peng YF, Wang F, Guo YY (2006) Evaluation of transgenic rice expressing Cry1Ac and CpTI against *Chilo suppressalis* and intrapopulation variation in susceptibility to Cry1Ac. *Environ Entomol* 35:1453–1459
108. Grainnet (2007) Monsanto and Dow Agrosiences launch “SmartStax,” industry’s first-ever eight-gene stacked combination in corn. http://www.grainnet.com/articles/Monsanto_and_Dow_Agrosiences_Launch_SmartStax_Industry_s_First_Ever_Eight_Gene_Stacked_Combination_in_Corn_-48374.html
109. Sanchis V, Agaisse H, Chauvaux J, Lereclus D (1996) Construction of new insecticidal *Bacillus thuringiensis* recombinant strains by using the sporulation non-dependent expression system of cryIIIA and a site specific recombination vector. *J Biotechnol* 48:81–96
110. Bosch D, Schipper B, van der Kleij H, de Maagd R, Stiekema W (1994) Recombinant *Bacillus thuringiensis* crystal proteins with new properties: possibilities for resistance management. *Bio/Technology* 12:915–919
111. de Maagd RA, Kwa MSG, van der Kleij H, Yamamoto T, Schipper B, Vlak JM, Stiekema WJ, Bosch D (1996) Domain III substitution in *Bacillus thuringiensis* delta-endotoxin Cry1A(b) results in superior toxicity for *Spodoptera exigua* and altered membrane protein recognition. *Appl Environ Microbiol* 62:1537–1543
112. de Maagd RA, Weemen-Hendriks M, Stiekema W, Bosch D (2000) *Bacillus thuringiensis* delta-endotoxin Cry1C domain III can function as a specificity determinant for *Spodoptera exigua* in different, but not all, Cry1-Cry1C hybrids. *Appl Environ Microbiol* 66:1559–1563
113. Rang C, Vachon V, Coux F, Carret C, Moar WJ, Brousseau R, Schwartz JL, Laprade R, Frutos R (2001) Exchange of domain I from *Bacillus thuringiensis* Cry1 toxins influences protoxin stability and crystal formation. *Curr Microbiol* 43:1–6
114. Naimov S, Dukiandjiev S, de Maagd RA (2003) A hybrid *Bacillus thuringiensis* delta-endotoxin gives resistance against a coleopteran and a lepidopteran pest in transgenic potato. *Plant Biotechnol J* 1:51–57
115. Dean DH, Rajamohan F, Lee MK, Wu SJ, Chen XJ, Alcantara E, Hussain SR (1996) Probing the mechanism of action of *Bacillus thuringiensis* insecticidal proteins by site-directed mutagenesis: a minireview. *Gene* 179:111–117
116. Rajamohan F, Alzate O, Cotrill JA, Curtiss A, Dean DH (1996) Protein engineering of *Bacillus thuringiensis* delta-endotoxin: mutations at domain II of CryIAb enhance receptor affinity and toxicity toward gypsy moth larvae. *Proc Natl Acad Sci USA* 93:14338–14343
117. Wu SJ, Koller CN, Miller DL, Bauer LS, Dean DH (2000) Enhanced toxicity of *Bacillus thuringiensis* Cry3A delta-endotoxin in coleopterans by mutagenesis in a receptor binding loop. *FEBS Lett* 473:227–232
118. Abdullah MAF, Alzate O, Mohammad M, McNall RJ, Adang MJ, Dean DH (2003) Introduction of *Culex* toxicity into *Bacillus thuringiensis* Cry4Ba by protein engineering. *Appl Environ Microbiol* 69:5343–5353

119. Tuntitippawan T, Boonserm P, Katzenmeier G, Angsuthanasombat C (2005) Targeted mutagenesis of loop residues in the receptor-binding domain of the *Bacillus thuringiensis* Cry4Ba toxin affects larvicidal activity. *FEMS Microbiol Lett* 242:325–332
120. Abdullah MAF, Dean DH (2004) Enhancement of Cry19Aa mosquitocidal activity against *Aedes aegypti* by mutations in the putative loop regions of domain II. *Appl Environ Microbiol* 70:3769–3771
121. Liu XS, Dean DH (2006) Redesigning *Bacillus thuringiensis* Cry1Aa toxin into a mosquito toxin. *Protein Eng Des Sel* 19:107–111
122. Ishikawa H, Hoshino Y, Motoki Y, Kawahara T, Kitajima M, Kitami M, Watanabe A, Bravo A, Soberon M, Honda A, Yaoi K, Sato R (2007) A system for the directed evolution of the insecticidal protein from *Bacillus thuringiensis*. *Mol Biotechnol* 36:90–101
123. Chandra A, Ghosh P, Mandaokar AD, Bera AK, Sharma RP, Das S, Kumar PA (1999) Amino acid substitution in alpha-helix 7 of Cry1Ac delta-endotoxin of *Bacillus thuringiensis* leads to enhanced toxicity to *Helicoverpa armigera* Hubner. *FEBS Lett* 458:175–179
124. Rupar MJ, Donovan WP, Groat RG, Slaney AC, Mattison JW, Johnson TB, Charles JF, Dumanior VC, DeBarjac H (1991) Two novel strains of *Bacillus thuringiensis* toxic to coleopterans. *Appl Environ Microbiol* 57:3337–3344
125. Bohorova N, Frutos R, Royer M, Estanol P, Pacheco M, Rascon Q, McLean S, Hoisington D (2001) Novel synthetic *Bacillus thuringiensis* cry1B gene and the cry1B-cry1Ab translational fusion confer resistance to southwestern corn borer, sugarcane borer and fall armyworm in transgenic tropical maize. *Theor Appl Genet* 103:817–826
126. Ho NH, Baisakh N, Oliva N, Datta K, Frutos R, Datta SK (2006) Translational fusion hybrid Bt genes confer resistance against yellow stem borer in transgenic elite vietnamese rice (*Oryza sativa* L.) cultivars. *Crop Sci* 46:781–789
127. Mehlo L, Gahakwa D, Nghia PT, Loc NT, Capell T, Gatehouse JA, Gatehouse AMR, Christou P (2005) An alternative strategy for sustainable pest resistance in genetically enhanced crops. *Proc Natl Acad Sci USA* 102:7812–7816
128. Gatehouse JA (2002) Plant resistance towards insect herbivores: a dynamic interaction. *New Phytol* 156:145–169
129. Barbosa JARG, Silva LP, Teles RCL, Esteves GF, Azevedo RB, Ventura MM, Freitas SM (2007) Crystal Structure of the Bowman-Birk inhibitor from *Vigna unguiculata* seeds in complex with beta-trypsin at 1.55 Å resolution and its structural properties in association with proteinases. *Biophys J* 92: 1638–1650
130. Garcia-Olmedo F, Salmedo G, Sanchez-Monge R, Gomez L, Royo J, Carbonero P (1987) Plant proteinaceous inhibitors of proteinases and α -amylases. *Oxf Surv Plant Mol Cell Biol* 4: 275–334
131. Ryan CA (1990) Protease inhibitors in plants – genes for improving defenses against insects and pathogens. *Annu Rev Phytopathol* 28:425–449
132. Orozco-Cardenas M, McGurl B, Ryan CA (1993) Expression of an antisense prosystemin gene in tomato plants reduces resistance toward *Manduca sexta* larvae. *Proc Natl Acad Sci USA* 90:8273–8276
133. Kessler A, Baldwin IT (2002) Plant responses to insect herbivory: the emerging molecular analysis. *Annu Rev Plant Biol* 53:299–328
134. Hilder VA, Gatehouse AMR, Sheerman SE, Barker RF, Boulter D (1987) A novel mechanism of insect resistance engineered into tobacco. *Nature* 330:160–163
135. Johnson R, Narvaez J, An G, Ryan C (1989) Expression of proteinase inhibitors I and II in transgenic tobacco plants: effects on natural defense against *Manduca sexta* larvae. *Proc Natl Acad Sci USA* 86:9871–9875
136. McManus MT, White DWR, McGregor PG (1994) Accumulation of a chymotrypsin inhibitor in transgenic tobacco can affect the growth of insect pests. *Transgenic Res* 3:50–58
137. Duan X, Li X, Xue Q, Abo-El-Saad M, Xu D, Wu R (1996) Transgenic rice plants harboring an introduced potato proteinase inhibitor II gene are insect resistant. *Nat Biotechnol* 14:494–496
138. Xu DP, Xue QZ, McElroy D, Mawal Y, Hilder VA, Wu R (1996) Constitutive expression of a cowpea trypsin-inhibitor gene, CpTI, in transgenic rice plants confers resistance to 2 major rice insect pests. *Mol Breed* 2:167–173
139. McGurl B, Orozco-Cardenas M, Pearce G, Ryan CA (1994) Overexpression of the prosystemin gene in transgenic tomato plants generates a systemic signal that constitutively induces proteinase inhibitor synthesis. *Proc Natl Acad Sci USA* 91:9799–9802
140. Ren F, Lu Y-T (2006) Overexpression of tobacco hydroxyproline-rich glycopeptide systemin precursor A gene in transgenic tobacco enhances resistance against *Helicoverpa armigera* larvae. *Plant Sci* 171:286–292
141. Bolter CJ, Jongsma MA (1995) Colorado potato beetles (*Leptinotarsa decemlineata*) adapt to proteinase inhibitors induced in potato leaves by methyl jasmonate. *J Insect Physiol* 41:1071–1078
142. Jongsma MA, Bakker PL, Peters J, Bosch D, Stiekma WJ (1995) Adaptations of *Spodoptera exigua* larvae to plant proteinase inhibitors by induction of gut proteinase activity insensitive to inhibition. *Proc Natl Acad Sci USA* 92:8041–8045
143. Harsulkar AM, Giri AP, Patankar AG, Gupta VS, Sainani MN, Ranjekar PK, Deshpande VV (1999) Successive use of non-host plant proteinase inhibitors required for effective inhibition of *Helicoverpa armigera* gut proteinases and larval growth. *Plant Physiol* 121:497–506
144. Bown DP, Wilkinson HS, Gatehouse JA (1997) Differentially regulated inhibitor-sensitive and insensitive protease genes from the phytophagous insect pest, *Helicoverpa armigera*, are members of complex multigene families. *Insect Biochem Mol Biol* 27:625–638
145. De Leo F, Bonade-Bottino MA, Ceci LR, Gallerani R, Jouanin L (1998) Opposite effects on *Spodoptera littoralis* larvae of high

- expression level of a trypsin proteinase inhibitor in transgenic plants. *Plant Physiol* 118:997–1004
146. Lepié JC, Bonade-Bottino M, Augustin S, Pilate G, Letan VD, Delplanque A, Cornu D, Jouanin L (1995) Toxicity to *Chrysomela tremulae* (Coleoptera, Chrysomelidae) of transgenic poplars expressing a cysteine proteinase inhibitor. *Mol Breed* 1:319–328
 147. Lecardonnel A, Chauvin L, Jouanin L, Beaujean A, Prevost G, Sangwan-Norree B (1999) Effects of rice cystatin I expression in transgenic potato on Colorado potato beetle larvae. *Plant Sci* 140:71–79
 148. Outchkourov NS, de Kogel WJ, Schuurman-de Bruin A, Abrahamson M, Jongsma MA (2004) Specific cysteine protease inhibitors act as deterrents of western flower thrips, *Frankliniella occidentalis* (Pergande), in transgenic potato. *Plant Biotechnol J* 2:439–448
 149. Outchkourov NS, de Kogel WJ, Wiegiers GL, Abrahamson M, Jongsma MA (2004) Engineered multidomain cysteine protease inhibitors yield resistance against western flower thrips (*Frankliniella occidentalis*) in greenhouse trials. *Plant Biotechnol J* 2:449–458
 150. Outchkourov NS, Rogelj B, Strukelj B, Jongsma MA (2003) Expression of sea anemone equistatin in potato: effects of plant proteases on heterologous protein production. *Plant Physiol* 133:379–390
 151. Abdeen A, Virgos A, Olivella E, Villanueva J, Aviles X, Gabarra R, Prat S (2005) Multiple insect resistance in transgenic tomato plants over-expressing two families of plant proteinase inhibitors. *Plant Mol Biol* 57:189–202
 152. Franco OL, Rigden DJ, Melo FR, Grossi-de-Sa MF (2002) Plant alpha-amylase inhibitors and their interaction with insect alpha-amylases - structure, function and potential for crop protection. *Eur J Biochem* 269:397–412
 153. Suzuki K, Ishimoto M, Kikuchi F, Kitamura K (1993) Growth-inhibitory effect of an alpha-amylase inhibitor from the wild common bean resistant to the mexican bean weevil (*Zabrotes subfasciatus*). *Jpn J Breed* 43:257–265
 154. Ishimoto M, Kitamura K (1991) Effect of absence of seed alpha-amylase inhibitor on the growth inhibitory activity to azuki bean weevil (*Callosobruchus chinensis*) in common bean (*Phaseolus vulgaris* L.). *Jpn J Breed* 41:231–240
 155. Moreno J, Chrispeels MJ (1989) A lectin gene encodes the α -amylase inhibitor of common bean. *Proc Natl Acad Sci USA* 86:7885–7889
 156. Nahoum V, Farisei F, Le-Berre-Anton V, Egloff MP, Rouge P, Poirio E, Payan F (1999) A plant-seed inhibitor of two classes of alpha-amylases: X-ray analysis of *Tenebrio molitor* larvae alpha-amylase in complex with the bean *Phaseolus vulgaris* inhibitor. *Acta Crystallogr Sect D* 55:360–362
 157. Silva CP, Terra WR, de Sa MFG, Samuels RI, Isejima EM, Bifano TD, Almeida JS (2001) Induction of digestive alpha-amylases in larvae of *Zabrotes subfasciatus* (Coleoptera: Bruchidae) in response to ingestion of common bean alpha-amylase inhibitor 1. *J Insect Physiol* 47:1283–1290
 158. Shade RE, Schroeder HE, Pueyo JJ, Tabe LM, Murdock LL, Higgins TJV, Chrispeels MJ (1994) Transgenic pea seeds expressing the alpha-amylase inhibitor of the common bean are resistant to bruchid beetles. *Bio/Technology* 12: 793–796
 159. Schroeder HE, Gollasch S, Moore A, Tabe LM, Craig S, Hardie DC, Chrispeels MJ, Spencer D, Higgins TJV (1995) Bean alpha-amylase inhibitor confers resistance to the pea weevil (*Bruchus pisorum*) in transgenic peas (*Pisum sativum* L.). *Plant Physiol* 107:1233–1239
 160. Ishimoto M, Sato T, Chrispeels MJ, Kitamura K (1996) Bruchid resistance of transgenic azuki bean expressing seed alpha-amylase inhibitor of common bean. *Entomol Exp Appl* 79: 309–315
 161. Morton RL, Schroeder HE, Bateman KS, Chrispeels MJ, Armstrong E, Higgins TJV (2000) Bean alpha-amylase inhibitor 1 in transgenic peas (*Pisum sativum*) provides complete protection from pea weevil (*Bruchus pisorum*) under field conditions. *Proc Natl Acad Sci USA* 97:3820–3825
 162. Sarmah BK, Moore A, Tate W, Molvig L, Morton RL, Rees DP, Chiaiese P, Chrispeels MJ, Tabe LM, Higgins TJV (2004) Transgenic chickpea seeds expressing high levels of a bean alpha-amylase inhibitor. *Mol Breed* 14:73–82
 163. Prescott VE, Campbell PM, Moore A, Mattes J, Rothenberg ME, Foster PS, Higgins TJV, Hogan SP (2005) Transgenic expression of bean alpha-amylase inhibitor in peas results in altered structure and immunogenicity. *J Agric Food Chem* 53: 9023–9030
 164. Collins CL, Eason PJ, Dunshea FR, Higgins TJV, King RH (2006) Starch but not protein digestibility is altered in pigs fed transgenic peas containing alpha-amylase inhibitor. *J Sci Food Agric* 86:1894–1899
 165. Li XH, Higgins TJV, Bryden WL (2006) Biological response of broiler chickens fed peas (*Pisum sativum* L.) expressing the bean (*Phaseolus vulgaris* L.) alpha-amylase inhibitor transgene. *J Sci Food Agric* 86:1900–1907
 166. Peumans WJ, van Damme EJM (1995) Lectins as plant defence proteins. *Plant Physiol* 109:347–352
 167. van Damme EJM, Peumans WJ, Barre A, Rougé P (1998) Plant lectins: a composite of several distinct families of structurally and evolutionary related proteins with diverse biological roles. *Crit Rev Plant Sci* 17:575–692
 168. Czalpa TH, Lang BA (1990) Effect of plant lectins on the larval development of European corn borer (Lepidoptera: Pyralidae) and southern corn rootworm (Coleoptera: Chrysomelidae). *J Econ Entomol* 83:2480–2485
 169. Murdock LL, Huesing JE, Nielsen SS, Pratt RC, Shade RE (1990) Biological effects of plant lectins on the cowpea weevil. *Phytochemistry* 29:85–89
 170. Fitches E, Gatehouse AMR, Gatehouse JA (1997) Effects of snowdrop lectin (GNA) delivered via artificial diet and transgenic plants on the development of tomato moth (*Lacanobia oleracea*) larvae in laboratory and glasshouse trials. *J Insect Physiol* 43:727–739

171. Boulter D, Edwards GA, Gatehouse AMR, Gatehouse JA, Hilder VA (1990) Additive protective effects of incorporating two different higher plant derived insect resistance genes in transgenic tobacco plants. *Crop Prot* 9:351–354
172. Gatehouse AMR, Davison GM, Newell CA, Merryweather A, Hamilton WDO, Burgess EPJ, Gilbert RJC, Gatehouse JA (1997) Transgenic potato plants with enhanced resistance to the tomato moth, *Lacanobia oleracea*: growth room trials. *Mol Breed* 3:49–63
173. Powell KS, Gatehouse AMR, Hilder VA, Gatehouse JA (1993) Antimetabolic effects of plant lectins and plant and fungal enzymes on the nymphal stages of two important rice pests, *Nilaparvata lugens* and *Nephotettix cinciteps*. *Entomol Exp Appl* 66:119–126
174. Powell KS, Gatehouse AMR, Hilder VA, van Damme EJM, Peumans WJ, Boonjawat J, Horsham K, Gatehouse JA (1995) Different antimetabolic effects of related lectins towards nymphal stages of *Nilaparvata lugens*. *Entomol Exp Appl* 75:61–65
175. Rao KV, Rathore KS, Hodges TK, Fu X, Stoger E, Sudhakar D, Williams S, Christou P, Bharathi M, Bown DP, Powell KS, Spence J, Gatehouse AMR, Gatehouse JA (1998) Expression of snowdrop lectin (GNA) in transgenic rice plants confers resistance to rice brown planthopper. *Plant J* 15:469–477
176. Nagadhara D, Ramesh S, Pasalu IC, Kondala Rao Y, Krishnaiah NV, Sarma NP, Bown DP, Gatehouse JA, Reddy VD, Rao KV (2003) Transgenic indica rice resistant to sap-sucking insects. *Plant Biotechnol* 1:231–240
177. Foissac X, Loc NT, Christou P, Gatehouse AMR, Gatehouse JA (2000) Resistance to green leafhopper (*Nephotettix virescens*) and brown planthopper (*Nilaparvata lugens*) in transgenic rice expressing snowdrop lectin (*Galanthus nivalis* agglutinin; GNA). *J Insect Physiol* 46:573–583
178. Nagadhara D, Ramesh S, Pasalu IC, Rao YK, Sarma NP, Reddy VD, Rao KV (2004) Transgenic rice plants expressing the snowdrop lectin gene (*gna*) exhibit high-level resistance to the whitebacked planthopper (*Sogatella furcifera*). *Theor Appl Genet* 109:1399–1405
179. Loc NT, Tinjuangjun P, Gatehouse AMR, Christou P, Gatehouse JA (2002) Linear transgene constructs lacking vector backbone sequences generate transgenic rice plants which accumulate higher levels of proteins conferring insect resistance. *Mol Breed* 9:231–244
180. Poulsen M, Kroghsbo S, Schroder M, Wilcks A, Jacobsen H, Miller A, Frenzel T, Danier J, Rychlik M, Shu QY, Emami K, Sudhakar D, Gatehouse A, Engel KH, Knudsen I (2007) A 90-day safety study in Wistar rats fed genetically modified rice expressing snowdrop lectin *Galanthus nivalis* (GNA). *Food Chem Toxicol* 45:350–363
181. Gatehouse AMR, Down RE, Powell KS, Sauvion N, Rahbe Y, Newell CA, Merryweather A, Hamilton WDO, Gatehouse JA (1996) Transgenic potato plants with enhanced resistance to the peach-potato aphid *Myzus persicae*. *Entomol Exp Appl* 79:295–307
182. Wang ZY, Zhang KW, Sun XF, Tang KX, Zhang JR (2005) Enhancement of resistance to aphids by introducing the snowdrop lectin gene *gna* into maize plants. *J Biosci* 30: 627–638
183. Bandyopadhyay S, Roy A, Das S (2001) Binding of garlic (*Allium sativum*) leaf lectin to the gut receptors of homopteran pests is correlated to its insecticidal activity. *Plant Sci* 161:1025–1033
184. Dutta I, Saha P, Majumder P, Sarkar A, Chakraborti D, Banerjee S, Das S (2005) The efficacy of a novel insecticidal protein, *Allium sativum* leaf lectin (ASAL), against homopteran insects monitored in transgenic tobacco. *Plant Biotechnol J* 3:601–611
185. Dutta I, Majumder P, Saha P, Ray K, Das S (2005) Constitutive and phloem specific expression of *Allium sativum* leaf agglutinin (ASAL) to engineer aphid (*Lipaphis erysimi*) resistance in transgenic Indian mustard (*Brassica juncea*). *Plant Sci* 169: 996–1007
186. Saha P, Majumder P, Dutta I, Ray T, Roy SC, Das S (2006) Transgenic rice expressing *Allium sativum* leaf lectin with enhanced resistance against sap-sucking insect pests. *Planta* 223:1329–1343
187. Saha P, Chakraborti D, Sarkar A, Dutta I, Basu D, Das S (2007) Characterization of vascular-specific *RSs1* and *rolC* promoters for their utilization in engineering plants to develop resistance against hemipteran insect pests. *Planta* 226:429–442
188. Saha P, Dasgupta I, Das S (2006) A novel approach for developing resistance in rice against phloem limited viruses by antagonizing the phloem feeding hemipteran vectors. *Plant Mol Biol* 62:735–752
189. Ryan CA (2000) The system in signaling pathway: differential activation of plant defensive genes. *Biochim Biophys Acta* 1477:112–121
190. Felton GW, Donato KK, Broadway RM, Duffey SS (1992) Impact of oxidized plant phenolics on the nutritional quality of dietary protein to a noctuid herbivore, *Spodoptera exigua*. *J Insect Physiol* 38:277–285
191. Melo GA, Shimizu MM, Mazzafera P (2006) Polyphenoloxidase activity in coffee leaves and its role in resistance against the coffee leaf miner and coffee leaf rust. *Phytochemistry* 67: 277–285
192. Wang JH, Constabel CP (2004) Polyphenol oxidase overexpression in transgenic *Populus* enhances resistance to herbivory by forest tent caterpillar (*Malacosoma disstria*). *Planta* 220:87–96
193. Dowd PF, Lagrimini LM (1997) Examination of different tobacco (*Nicotiana* spp.) types under- and overproducing tobacco anionic peroxidase for their leaf resistance to *Helicoverpa zea*. *J Chem Ecol* 23:2357–2370
194. Dowd PF, Lagrimini LM (2006) Examination of the biological effects of high anionic peroxidase production in tobacco plants grown under field conditions. I. Insect pest damage. *Transgenic Res* 15:197–204
195. Dowd PF, Zuo WN, Gillikin JW, Johnson ET, Boston RS (2003) Enhanced resistance to *Helicoverpa zea* in tobacco

- expressing an activated form of maize ribosome-inactivating protein. *J Agric Food Chem* 51:3568–3574
196. Lawrence SD, Novak NG (2006) Expression of poplar chitinase in tomato leads to inhibition of development in Colorado potato beetle. *Biotechnol Lett* 28:593–599
 197. Dowd PF, Johnson ET, Pinkerton TS (2007) Oral toxicity of beta-N-acetyl hexosaminidase to insects. *J Agric Food Chem* 55:3421–3428
 198. Fitches E, Wilkinson H, Bell H, Bown DP, Gatehouse JA, Edwards JP (2004) Cloning, expression and functional characterisation of chitinase from larvae of tomato moth (*Lacanobia oleracea*): a demonstration of the insecticidal activity of insect chitinase. *Insect Biochem Mol Biol* 34:1037–1050
 199. Ding XF, Gopalakrishnan B, Johnson LB, White FF, Wang XR, Morgan TD, Kramer KJ, Muthukrishnan S (1998) Insect resistance of transgenic tobacco expressing an insect chitinase gene. *Transgenic Res* 7:77–84
 200. Saguez J, Hainez R, Cherqui A, Van Wuytswinkel O, Jeanpierre H, Lebon G, Noiraud N, Beaujean A, Jouanin L, Laberche JC, Vincent C, Giordanengo P (2005) Unexpected effects of chitinases on the peach-potato aphid (*Myzus persicae* Sulzer) when delivered via transgenic potato plants (*Solanum tuberosum* Linne) and *in vitro*. *Transgenic Res* 14:57–67
 201. Corrado G, Arciello S, Fanti P, Fiandra L, Garonna A, Digilio MC, Lorito M, Giordana B, Pennacchio F, Rao R (2007) The Chitinase A from the baculovirus AcMNPV enhances resistance to both fungi and herbivorous pests in tobacco. *Transgenic Res* 17:557–571. (published online: DOI: 10.1007/s11248-007-9129-4)
 202. Wittstock U, Gershenzon J (2002) Constitutive plant toxins and their role in defense against herbivores and pathogens. *Curr Opin Plant Biol* 5:300–307
 203. Tattersall DB, Bak S, Jones PR, Olsen CE, Nielsen JK, Hansen ML, Hoj PB, Møller BL (2001) Resistance to an herbivore through engineered cyanogenic glucoside synthesis. *Science* 293:1826–1828
 204. Bak S, Olsen CE, Halkier BA, Møller BL (2000) Transgenic tobacco and *Arabidopsis* plants expressing the two multifunctional sorghum cytochrome P450 enzymes, CYP79A1 and CYP71E1, are cyanogenic and accumulate metabolites derived from intermediates in dhurrin biosynthesis. *Plant Physiol* 123:1437–1448
 205. Kristensen C, Morant M, Olsen CE, Ekstrom CT, Galbraith DW, Møller BL, Bak S (2005) Metabolic engineering of dhurrin in transgenic *Arabidopsis* plants with marginal inadvertent effects on the metabolome and transcriptome. *Proc Natl Acad Sci USA* 102:1779–1784
 206. Ogunlabi OO, Agboola FK (2007) A soluble beta-cyanoalanine synthase from the gut of the variegated grasshopper *Zonocerus variegatus* (L.). *Insect Biochem Mol Biol* 37:72–79
 207. Mikkelsen MD, Halkier BA (2003) Metabolic engineering of valine- and isoleucine-derived glucosinolates in *Arabidopsis* expressing CYP79D2 from cassava. *Plant Physiol* 131:773–779
 208. Franks TK, Powell KS, Choimes S, Marsh E, Iocco P, Sinclair BJ, Ford CM, van Heeswijk R (2006) Consequences of transferring three sorghum genes for secondary metabolite (cyanogenic glucoside) biosynthesis to grapevine hairy roots. *Transgenic Res* 15:181–195
 209. Kim YS, Uefuji H, Ogita S, Sano H (2006) Transgenic tobacco plants producing caffeine: a potential new strategy for insect pest control. *Transgenic Res* 15:667–672
 210. Wei S, Semel Y, Bravdo BA, Czosnek H, Shoseyov O (2007) Expression and subcellular compartmentation of *Aspergillus niger* beta-glucosidase in transgenic tobacco result in an increased insecticidal activity on whiteflies (*Bemisia tabaci*). *Plant Sci* 172:1175–1181
 211. Aharoni A, Jongsma MA, Bouwmeester HJ (2005) Volatile science? Metabolic engineering of terpenoids in plants. *Trends Plant Sci* 10:594–602
 212. Wang E, Wang R, DeParasis J, Loughrin JH, Gan S, Wagner GJ (2001) Suppression of a P450 hydroxylase gene in plant trichome glands enhances natural-product-based aphid resistance. *Nat Biotechnol* 19:371–374
 213. Aharoni A, Giri AP, Deuerlein S, Griepink F, de Kogel WJ, Verstappen FWA, Verhoeven HA, Jongsma MA, Schwab W, Bouwmeester HJ (2003) Terpenoid metabolism in wild-type and transgenic *Arabidopsis* plants. *Plant Cell* 15:2866–2884
 214. Kappers IF, Aharoni A, van Herpen TWJM, Luckerhoff LLP, Dicke M, Bouwmeester HJ (2005) Genetic engineering of terpenoid metabolism attracts bodyguards to *Arabidopsis*. *Science* 309:2070–2072
 215. Schnee C, Kollner TG, Held M, Turlings TCJ, Gershenzon J, Degenhardt J (2006) The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores. *Proc Natl Acad Sci USA* 103:1129–1134
 216. Beale MH, Birkett MA, Bruce TJA, Chamberlain K, Field LM, Huttly AK, Martin JL, Parker R, Phillips AL, Pickett JA, Prosser IM, Shewry PR, Smart LE, Wadhams LJ, Woodcock CM, Zhang YH (2006) Aphid alarm pheromone produced by transgenic plants affects aphid and parasitoid behavior. *Proc Natl Acad Sci USA* 103:10509–10513
 217. Johnson ET, Dowd PF (2004) Differentially enhanced insect resistance, at a cost, in *Arabidopsis thaliana* constitutively expressing a transcription factor of defensive metabolites. *J Agric Food Chem* 52:5135–5138
 218. Johnson ET, Berhow MA, Dowd PF (2007) Expression of a maize Myb transcription factor driven by a putative silk-specific promoter significantly enhances resistance to *Helicoverpa zea* in transgenic maize. *J Agric Food Chem* 55:2998–3003
 219. Maiti IB, Dey N, Pattanaik S, Dahlman DL, Rana RL, Webb BA (2003) Antibiosis-type insect resistance in transgenic plants expressing a teratocyte secretory protein (TSP14) gene from a hymenopteran endoparasite (*Microplitis croceipes*). *Plant Biotechnol J* 1:209–219
 220. Tortiglione C, Fogliano V, Ferracane R, Fanti P, Pennacchio F, Monti LM, Rao R (2003) An insect peptide engineered into the

- tomato prosystemin gene is released in transgenic tobacco plants and exerts biological activity. *Plant Mol Biol* 53:891–902
221. Bowen DJ, Ensign JC (1998) Purification and characterization of a high-molecular-weight insecticidal protein complex produced by the entomopathogenic bacterium *Photobacterium luminescens*. *Appl Environ Microbiol* 64:3029–3035
 222. Bowen D, Rocheleau TA, Blackburn M, Andreev O, Golubeva E, Bhartia R, French-Constant RH (1998) Insecticidal toxins from the bacterium *Photobacterium luminescens*. *Science* 280: 2129–2132
 223. French-Constant RH, Dowling A, Waterfield NR (2007) Insecticidal toxins from *Photobacterium* bacteria and their potential use in agriculture. *Toxicon* 49:436–451
 224. Guo LN, Fatig RO, Orr GL, Schafer BW, Strickland JA, Sukhapinda K, Woodworth AT, Petell JK (1999) *Photobacterium luminescens* W-14 insecticidal activity consists of at least two similar but distinct proteins - purification and characterization of toxin A and toxin B. *J Biol Chem* 274:9836–9842
 225. Liu D, Burton S, Glancy T, Li ZS, Hampton R, Meade T, Merlo DJ (2003) Insect resistance conferred by 283-kDa *Photobacterium luminescens* protein TcdA in *Arabidopsis thaliana*. *Nat Biotechnol* 21:1222–1228
 226. Purcell JP, Greenplate JT, Jennings MG, Ryerse JS, Pershing JC, Sims SR, Prinsen MJ, Corbin DR, Tran M, Sammons RD, Stonard RJ (1993) Cholesterol oxidase - a potent insecticidal protein active against boll weevil larvae. *Biochem Biophys Res Commun* 196:1406–1413
 227. Shen Z, Corbin DR, Greenplate JT, Grebenok RJ, Galbraith DW, Purcell JP (1997) Studies on the mode of action of cholesterol oxidase on insect midgut membranes. *Arch Insect Biochem Physiol* 34:429–442
 228. Corbin DR, Grebenok RJ, Ohnmeiss TE, Greenplate JT, Purcell JP (2001) Expression and chloroplast targeting of cholesterol oxidase in transgenic tobacco plants. *Plant Physiol* 126:1116–1128
 229. Morgan TD, Oppert B, Czapl TH, Kramer KJ (1993) Avidin and streptavidin as insecticidal and growth-inhibiting dietary proteins. *Entomol Exp Appl* 69:97–108
 230. Markwick NP, Christeller JT, Docherty LC, Lilley CM (2001) Insecticidal activity of avidin and streptavidin against four species of pest Lepidoptera. *Entomol Exp Appl* 98:59–66
 231. Hood EE, Witcher DR, Maddock S, Meyer T, Baszczynski C, Bailey M, Flynn P, Register J, Marshall L, Bond D, Kulisek E, Kusnadi A, Evangelista R, Nikolov Z, Wooge C, Mehig R, Hernan R, Kappel WK, Ritland D, Li CP, Howard JA (1997) Commercial production of avidin from transgenic maize: characterization of transformant, production, processing, extraction and purification. *Mol Breed* 3:291–306
 232. Kramer KJ, Morgan TD, Throne JE, Dowell FE, Bailey M, Howard JA (2000) Transgenic avidin maize is resistant to storage insect pests. *Nat Biotechnol* 18:670–674
 233. Burgess EPJ, Malone LA, Christeller JT, Lester MT, Murray C, Philip BA, Phung MM, Tregidga EL (2002) Avidin expressed in transgenic tobacco leaves confers resistance to two noctuid pests, *Helicoverpa armigera* and *Spodoptera litura*. *Transgenic Res* 11:185–198
 234. Murray C, Sutherland PW, Phung MM, Lester MT, Marshall RK, Christeller JT (2002) Expression of biotin-binding proteins, avidin and streptavidin, in plant tissues using plant vacuolar targeting sequences. *Transgenic Res* 11:199–214
 235. Markwick NP, Docherty LC, Phung MM, Lester MT, Murray C, Yao JL, Mitra DS, Cohen D, Beuning LL, Kutty-Amma S, Christeller JT (2003) Transgenic tobacco and apple plants expressing biotin-binding proteins are resistant to two cosmopolitan insect pests, potato tuber moth and lightbrown apple moth, respectively. *Transgenic Res* 12:671–681
 236. Yoza K, Imamura T, Kramer KJ, Morgan TD, Nakamura S, Akiyama K, Kawasaki S, Takaiwa F, Ohtsubo K (2005) Avidin expressed in transgenic rice confers resistance to the stored-product insect pests *Tribolium confusum* and *Sitotroga cerealella*. *Biosci Biotechnol Biochem* 69:966–971
 237. Ginzberg I, Perl A, Genser M, Winer S, Nemas C, Kapulnik Y (2004) Expression of streptavidin in tomato resulted in abnormal plant development that could be restored by biotin application. *J Plant Physiol* 161:611–620
 238. Flinn PW, Kramer KJ, Throne JE, Morgan TD (2006) Protection of stored maize from insect pests using a two-component biological control method consisting of a hymenopteran parasitoid, *Theocolax elegans*, and transgenic avidin maize powder. *J Stored Prod Res* 42:218–225
 239. Cooper SG, Douches DS, Grafius EJ (2006) Insecticidal activity of avidin combined with genetically engineered and traditional host plant resistance against Colorado potato beetle (Coleoptera: Chrysomelidae) larvae. *J Econ Entomol* 99: 527–536
 240. Turner CT, Davy MW, MacDiarmid RM, Plummer KM, Birch NP, Newcomb RD (2006) RNA interference in the light brown apple moth, *Epiphyas postvittana* (Walker) induced by double-stranded RNA feeding. *Insect Mol Biol* 15:383–391
 241. Mao Y-B, Cai W-J, Wang J-W, Hong G-J, Tao X-Y, Wang L-J, Huang Y-P, Chen X-Y (2007) Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol. *Nat Biotechnol* 25:1307–1313
 242. Baum JA, Bogaert T, Clinton W, Heck GR, Feldmann P, Ilagan O, Johnson S, Plaetinck G, Munyikwa T, Pleau M, Vaughn T, Roberts J (2007) Control of coleopteran insect pests through RNA interference. *Nat Biotechnol* 25:1322–1326
 243. Benbrook C (2009) Impacts of genetically engineered crops on pesticide use in the United States: The first thirteen years. *The Organic Centre; Critical Issue Report*. http://www.organic-center.org/reportfiles/13Years20091126_FullReport.pdf
 244. Gatehouse AMR, Ferry N, Raemaekers RJM (2002) The case of the monarch butterfly: a verdict is returned. *Trends Genet* 18:249–251
 245. Tabashnik BE, Van Rensburg JBJ, Carriere Y (2009) Field-evolved insect resistance to *Bt* crops: Definition, theory, and data. *J Econ Entomol* 102:2011–2025

246. Nunez-Farfan J, Fornoni J, Luis Valverde P (2007) The evolution of resistance and tolerance to herbivores. *Annu Rev Ecol Evol Syst* 38:541–566

Books and reviews

- Carozzi N, Koziel M, eds. (1997) *Advances in Insect Control; The Role of Transgenic Plants*. Taylor and Francis, London
- Romeis J, Shelton AM, Kennedy G eds. (2008) *Integration of Insect-Resistant Genetically Modified Crops within IPM Programs (Progress in Biological Control)*. Springer
- Lemaux PG (2008) Genetically engineered plants and foods: a scientist's analysis of the issues (part I). *Annu Rev Plant Biol* 59:771–812
- Lemaux PG (2009) Genetically engineered plants and foods: a scientist's analysis of the issues (part II). *Annu Rev Plant Biol* 60:511–559
- Park JR, McFarlane I, Phipps RH, Ceddia G (2011) The role of transgenic crops in sustainable development. *Plant Biotech J* 9:2–21

GEN-IV Reactors

TAEK K. KIM

Nuclear Engineering Division, Argonne National Laboratory, Argonne, IL, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 GEN-IV Nuclear Systems
 Future Directions
 Bibliography

Glossary

- Breeding ratio** Ratio of fission atom production to fissile atom destruction during a certain time interval in a nuclear system.
- Closed fuel cycle (full recycle)** One of the nuclear fuel cycle options, in which all actinides in the used nuclear fuel are separated and recycled to reduce the radiotoxicity of a geological repository while enhancing uranium utilization.
- Energy sustainability** Ability to meet the energy needs of the present generation while enhancing the

ability of the future generation. In GEN-IV, the sustainability is measured by utilization of uranium resource without creating any weakness in economics and environmental goals.

GFR Gas-cooled Fast Reactor, which features a fast reactor and closed fuel cycle.

GIF Generation IV international forum, which is a cooperative international endeavor organized to carry out the R&D needed to establish the feasibility and performance capabilities of GEN-IV nuclear systems.

LFR Lead-cooled Fast Reactor, which features a fast reactor and closed fuel cycle.

MSR Molten Salt Reactor, which features thermal, epithermal, or fast reactor and closed fuel cycle.

Open fuel cycle (once-through cycle) One of the nuclear fuel cycle options, in which the used nuclear fuel discharged from a nuclear system is stored for some period of time and disposed in a geological repository isolating from environment.

Pyroprocessing The complete set of operations developed in USA. Integral Fast Reactor program based on the pyrometallurgical and electrochemical processes for recovering actinide elements from the used nuclear fuel and recycling them.

SCWR Supercritical Water Reactor, which features either thermal or fast reactor and open or closed fuel cycle.

SFR Sodium-cooled Fast Reactor, which features a fast reactor and closed fuel cycle.

Uranium utilization Ratio of uranium mass used in a nuclear system for energy generation to the uranium mass required by the nuclear system in a nuclear fuel cycle option.

VHTR Very-High-Temperature Reactor, which features a thermal reactor and open fuel cycle.

Definition of the Subject

Generation-IV reactors are a set of nuclear reactors currently being developed under international collaborations targeting sustainability, safety and reliability, high economics, proliferation resistance, and physical protection of nuclear energy. Nuclear systems have been developed over a number of decades and have evolved to the third generation from the first generation of prototypes constructed in 1950s and 1960s,

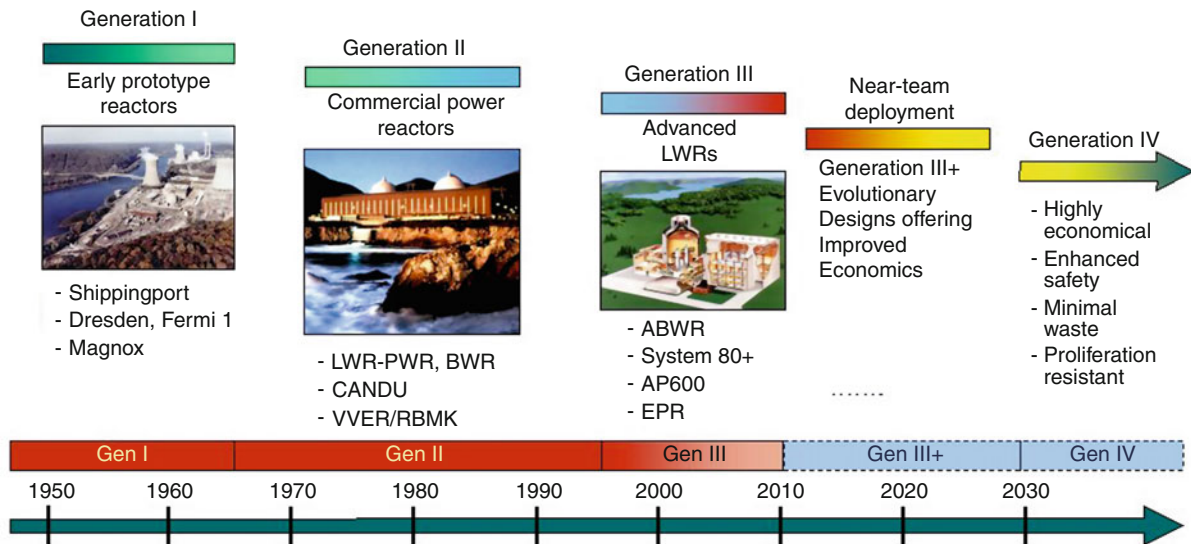
via the second generation of the commercial reactors operated worldwide after 1970s. While the third generation nuclear systems are currently proposed to the potential customers and under constructions with significant evolutionary in economics and safety based on lessons learnt through plenty reactor operations, nuclear experts from around the world began formulating the requirements for a generation IV of nuclear systems concerning over energy resource availability, climate change, air quality, and energy security. Six systems have been selected for further R&D as generation IV nuclear systems by Generation IV International Forum (GIF), which is a cooperative international endeavor organized to carry out the R&D needed to establish the feasibility and performance capabilities of Generation IV systems. The six systems are Gas-cooled Fast Reactor, Lead-cooled Fast Reactor, Molten Salt Reactor, Sodium-cooled Fast Reactor, Supercritical-Water Reactor, and Very-High-Temperature Reactor.

Introduction

Nuclear energy systems have evolved up to the third generation: a first generation of prototypes constructed in 1950 and 1960; a second generation of commercial nuclear power plants built from 1970, most of which

are in operation today; and a third generation of advanced nuclear reactors, called Generation III/III+, which incorporate technical progress based on lessons learnt through more than 10,000 reactor-years of operation. While the generation III/III+ nuclear systems are currently proposed to the potential customers and under constructions with significant evolutionary in economics and safety, nuclear experts from around the world indicated that further advances in nuclear energy systems are required to better meet the rapid growth of environment friendly, highly economic, and secure nuclear energy in both industrialized and developing countries. In particular, it is now globally recognized that the nuclear energy is the practically available massive energy source without greenhouse gas emission among numerous options. To meet these needs, the international nuclear community has engaged in a wide-range discussion on the development of next generation nuclear energy systems known as *Generation IV* (GEN-IV) targeting the deployment around 2030. [Figure 1](#) shows the evolution of the nuclear energy systems.

Nine countries, Argentina, Brazil, Canada, France, Japan, the Republic of Korea, the Republic of South Africa, the UK, and the USA, have initially joined together to form the Generation IV International



GEN-IV Reactors. Figure 1
Evolution of nuclear systems

Forum (GIF) [1] for developing GEN-IV nuclear systems that can be licensed, constructed, and operated in a manner that will provide competitively priced and reliable energy products while satisfactorily addressing nuclear safety, waste, proliferation, and public perception concerns. Now, the GIF consists of 13 membership countries added by China, Euratom, Russia, and Switzerland, and two permanent observers of International Atomic Energy Agency (IAEA) and the Organization for Economic Cooperation and Development Nuclear Energy Agency (OECD/NEA).

Beginning in 2000, more than 100 of nuclear experts from the countries constituting the GIF began to discuss for development of the GEN-IV technology roadmap in order to select the GEN-IV nuclear systems. As the first effort in the technology roadmap project [2], eight goals for the GEN-IV were defined in the four broad areas as shown in Table 1.

Since the eight goals are all equally important, the promising GEN-IV systems should ideally advance each and not create a weakness in one goal to gain strength in another. Under this central feature of the technical roadmap project, a series of GIF meeting was held in 2002 to conduct the selection process of the GEN-IV nuclear energy systems. The candidate systems were screened by the GIF expert group and six nuclear systems were selected on a consensus of the GIF membership countries such that the systems are the most promising and worthy of collaborative developments. The selected six systems for further R&D are alphabetically

- Gas-cooled Fast Reactor System (GFR),
- Lead-cooled Fast Reactor System (LFR),
- Molten Salt Reactor System (MSR),
- Sodium-cooled Fast Reactor System (SFR),
- Supercritical-water-cooled Reactor System (SCWR),
- Very-High-Temperature Reactor System (VHTR).

GEN-IV Nuclear Systems

In Table 2, the primary characteristics of the GEN-IV nuclear systems are summarized. In the roadmap project, it was recognized that the GIF countries would have perspectives on their priority missions for GEN-IV nuclear systems, which can be summarized as electricity generation, hydrogen production, and

GEN-IV Reactors. Table 1 Goal for generation IV nuclear energy systems

Sustainability 1	Generation IV nuclear energy systems will provide sustainable energy generation that meets clean air objectives and promotes long-term availability of systems and effective fuel utilization for worldwide energy production.
Sustainability 2	Generation IV nuclear energy systems will minimize and manage their nuclear waste and notably reduce the long-term stewardship burden, thereby improving protection for the public health and the environment.
Economics 1	Generation IV nuclear energy systems will have a clear life-cycle cost advantage over other energy sources.
Economics 2	Generation IV nuclear energy systems will have a level of financial risk comparable to other energy projects.
Safety and Reliability 1	Generation IV nuclear energy systems operations will excel in safety and reliability.
Safety and Reliability 2	Generation IV nuclear energy systems will have a very low likelihood and degree of reactor core damage.
Safety and Reliability 3	Generation IV nuclear energy systems will eliminate the need for off-site emergency response.
Proliferation Resistance and Physical Protection 1	Generation IV nuclear energy systems will increase the assurance that they are a very unattractive and the least desirable route for diversion or theft of weapons-usable materials, and provide increased physical protection against acts of terrorism.

GEN-IV Reactors. Table 2 Summary of GEN-IV nuclear systems

	Coolant	Neutron spectrum	Coolant exit temp. (°C)	Fuel cycle	Size (MWe)
GFR	Helium	Fast	850	Closed	1,200
LFR	Lead	Fast	480–800	Closed	50–1,200
MSR	Fluoride salt	Fast/thermal	700–800	Closed	1,000
SFR	Sodium	Fast	550	Closed	30–2,000
SCWR	Water	Thermal/fast	510–625	Open/closed	300–1,500
VHTR	Helium	Thermal	900–1,000	Open	250–300

high-level radioactive material management. All six GEN-IV nuclear systems have electricity applications, while the high temperature and fast neutron spectrum are required for the hydrogen generation and high-level radioactive material management, respectively. The high temperature systems such as VHTR, GFR, LFR, and MSR have potential applications in hydrogen production. By reprocessing and recycling of actinides, the fast reactor systems such as SFR, GFR, and LFR would provide a significant reduction in radiotoxicity of all wastes.

GFR – Gas-cooled Fast Reactor

The Gas-cooled Fast Reactor system features a fast-spectrum helium-cooled reactor and closed fuel cycle. Figure 2 shows the schematic of the GFR, which uses a direct-cycle helium turbine for electricity. Like thermal-spectrum helium-cooled reactors such as the Gas Turbine-Modular Helium Reactor (GT-MHR [3]) and the Pebble Bed Modular Reactor (PBMR [4]), the high outlet temperature of the helium coolant makes it possible to deliver not only electricity, but also process heat for hydrogen production with a high conversion efficiency. Through the combination of a fast-neutron spectrum and closed fuel cycle options, the GFR can manage the high-level radioactive waste isotopes.

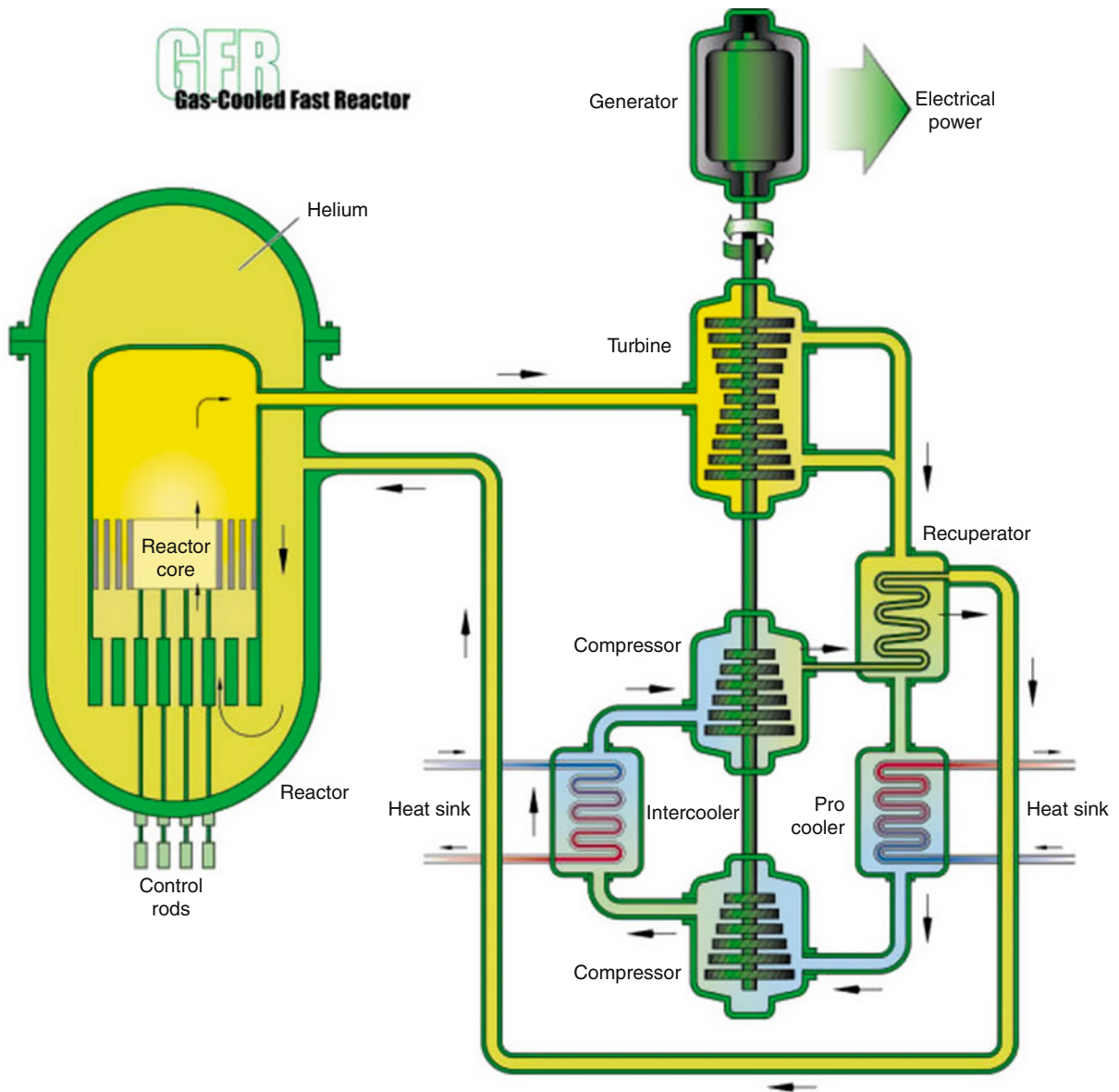
The technology base for the GFR includes a number of thermal-spectrum gas reactor plants, as well as a few fast-spectrum gas-cooled reactor designs. Past pilot and demonstration projects include decommissioned reactors such as the Dragon Project [5] built and operated in the UK, the AVR [6] and the Thorium High-Temperature Reactor (THTR [7]) built and operated in Germany, and Peach Bottom and Fort St Vrain [8] built

and operated in the USA. Ongoing demonstrations include the High-Temperature engineering Test Reactor (HTTR [9]) in Japan, which reached full power (30 MWth) using prismatic fuel compacts in 1999, and the High-Temperature Gas-cooled Reactor (HTR-10 [10]) in China, which reached 10 MWth in 2002 using pebble fuel.

A 300-MWth pebble bed modular demonstration plant is being designed by PBMR Pty for deployment in South Africa and a consortium of Russian institutes is designing a 300-MWth GT-MHR in cooperation with General Atomics. The design of the PBMR and GT-MHR reactor systems, fuel, and materials are evolutionary advances of the demonstrated technology, except for the Brayton-cycle helium turbine and implementation of modularity in the plant design. The GFR may benefit from development of these technologies, as well as development of innovative fuel and very-high-temperature materials for the VHTR.

Spent fuel treatment for the GFR can be accomplished with aqueous processes similar to those of the SFR but qualified for the unique GFR fuel form. A composite ceramic–ceramic fuel (CERCER) with closely packed, coated (U, Pu)C kernels or fibers is considered as the primary option for fuel development. Alternative fuel options for development include fuel particles with large (U, Pu)C kernels and thin coatings, or ceramic-clad, solid-solution metal (CERMET) fuels. The need for a high density of heavy metal elements in the fuel leads to actinide-carbides as the reference fuel and actinide-nitrides with 99.9% enriched nitrogen as the backup.

The reference material for the structure is reinforced ceramic comprising a silicon carbide composite matrix ceramic. The fuel compound is made of pellets of mixed



GEN-IV Reactors. Figure 2
Gas-cooled fast reactor

uranium-plutonium-minor actinide carbide. A leaktight barrier made of a refractory metal or of Si-based multi-layer ceramics is added to prevent fission products' diffusion through the clad.

Neither experimental reactors nor prototypes of the GFR system have been licensed or built; therefore, the construction and operation of a first experimental reactor – 50 MWth Experimental Technology Demonstration Reactor (ETDR [11]) – is proposed with an

extended performance phase to qualify key technologies. A technology demonstration reactor would qualify key technologies and could be put into operation by 2025.

Unlike the VHTR, which uses its considerable thermal mass to limit the rise of core temperature during transients, the GFR requires the development of a number of unique subsystems to provide defense in depth for its considerably higher power density core. These include

a robust decay heat removal system with added provisions for natural circulation heat removal, such as a low-pressure-drop core. The secondary circuit uses a He–N₂ gas mixture with an indirect combined (Brayton and bottoming steam) power cycle to achieve more than 45% thermal efficiency.

A gastight envelope acting as additional guard containment is provided to maintain a backup pressure in case of large gas leak from the primary system. It is a metallic vessel, initially filled with nitrogen slightly over the atmospheric pressure to reduce air ingress potential. This unique component limits the consequence of coincident first and second safety barrier rupture (i.e., the fuel cladding and the primary system). Dedicated loops for decay heat removal (in case of emergency) are directly connected to the primary circuit using cross duct piping from the pressure vessel and are equipped with heat exchangers and blowers.

Many of the structural materials and methods are being adopted from the VHTR, including the reactor pressure vessel, hot duct materials, and design approach. The pressure vessel is a thick metallic structure of martensitic chromium steel, ensuring negligible creep at operating temperature. The primary system is comprised of three main loops of 800 MWth, each fitted with compact intermediate heat exchangers and a gas blower enclosed in a single vessel.

As a high-temperature and high-power density system, the GFR gives special attention to safety and materials management for both economics and non-proliferation. During the viability phase that is underway now, there is special interest in examining the use of pin-type fuel with a small diameter, fuel and core performance optimized for a simplified GFR having no minor actinide recycle, but with limited Pu breeding and low fuel burnup, core outlet temperature optimized to balance efficiency with materials limits, and the potential of prestressed concrete vessel technology to replace the guard vessel.

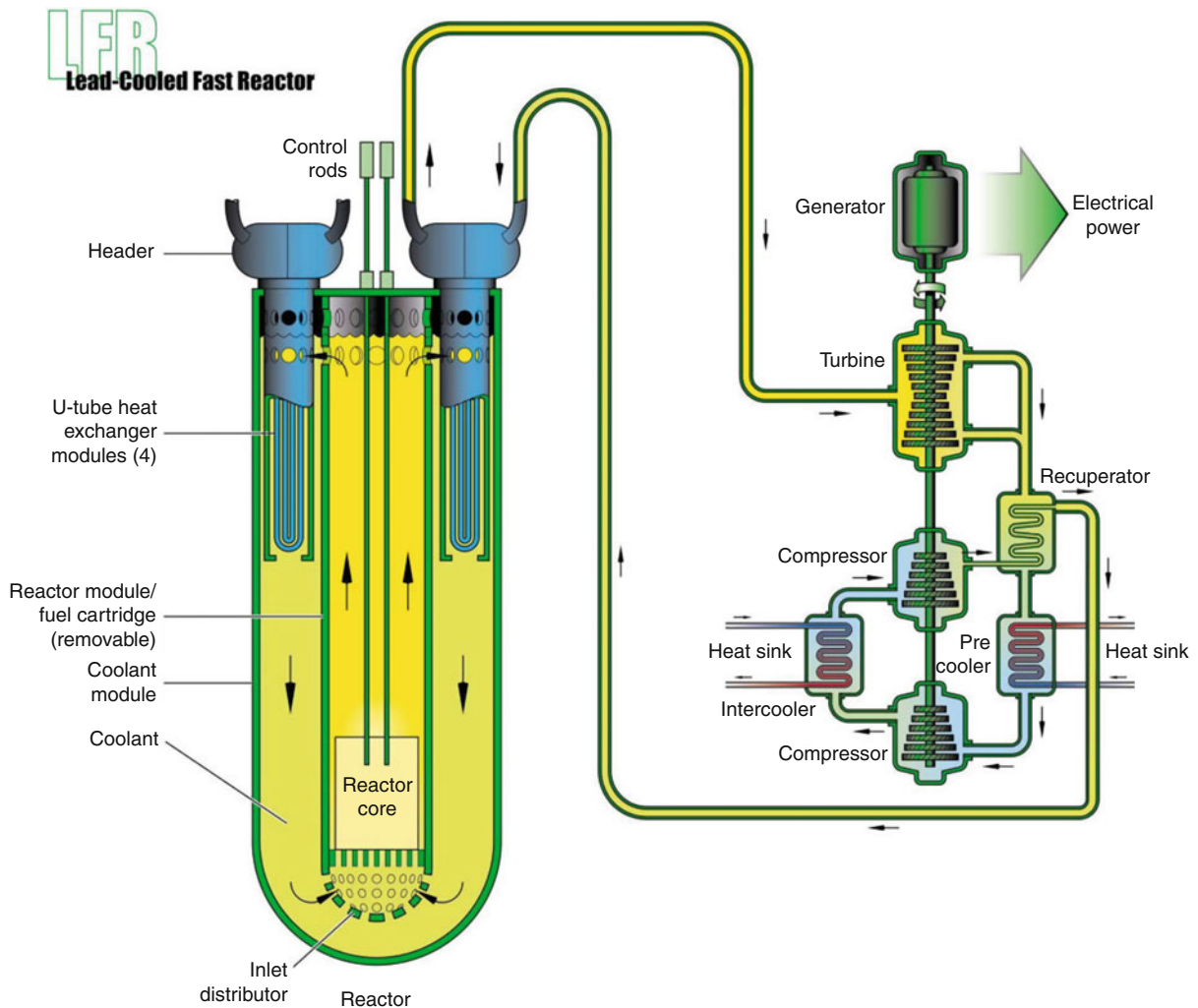
LFR – Lead-cooled Fast Reactor

The Lead-cooled Fast Reactor is similar to the sodium-cooled fast reactor in terms of neutron spectrum, fuel cycles, and the missions, but the coolant materials are changed to lead (Pb) or lead–bismuth (Pb–Bi). The lead coolant exhibits very low parasitic neutron

absorption in fast neutron spectral environment, and this enables the sustainability and fuel cycle benefits traditionally associated with SFR. However, lead does not react readily with air, water, or carbon dioxide, which can eliminate the concerns about vigorous exothermic reactions. It has a high boiling temperature. The need to operate under high pressure and the prospect of boiling or flashing in case of pressure reduction are eliminated. Figure 3 shows the schematic of the LFR.

There are several potentials for advances compared to state-of-the-art liquid metal fast reactors. Innovations in heat transport and energy conversion are a central feature of the LFR options. Innovations in heat transport are afforded by natural circulation, lift pumps, in-vessel steam generators, and other features. Innovations in energy conversion are afforded by rising to higher temperatures than liquid sodium allows, and by reaching beyond the traditional superheated Rankine cycle to supercritical Brayton cycle or process heat applications such as hydrogen production and desalination. The favorable neutronics of coolant enable low power density, natural circulation-cooled reactors with fissile self-sufficient core designs that maintain criticality over 15-year refueling interval. For modular and large units, more conventional higher power density, forced circulation, and shorter refueling intervals are used, but these units benefit from the improved heat transport and energy conversion technology. The favorable properties of lead coolant and nitride fuel, combined with high-temperature structural materials can extend the reactor coolant outlet temperature up to 800°C, which is potentially suitable for hydrogen manufacture and other process heat applications.

Two types of LFR reactors were used in Russian submarines of the 1970s with the 155 MWth LFR reactors, OK-550 and BM-440. Recently, Russian joint venture AKME Engineering announced to develop a commercial LFR called SBVR-100 [12]. The core is based on the former LFR reactors used in the submarines and will produce 100 MWe electricity from gross thermal power of 280 MWth, about twice that of the submarine reactors. The coolant is 495°C and 16.5% enriched uranium oxide fuel is used with the refueling schedule of 7–8 years. The small lead-cooled fast reactor concept known as the small secure transportable autonomous reactor (SSTAR [13]) has been under ongoing development as part of the US

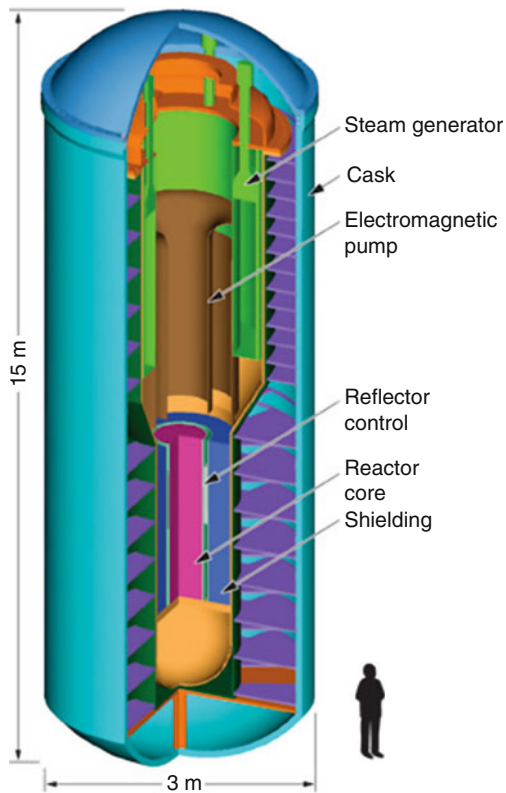


GEN-IV Reactors. Figure 3
Lead-cooled fast reactor

advanced nuclear energy systems programs (see Fig. 4). It is a system designed to provide energy security to developing nations while incorporating features to achieve nonproliferation goals. A 600 MWe European Lead-cooled system (ELSY [14]) has been under development since 2006. The ELSY project aims at the demonstration that it is possible to design a competitive and safe fast power reactor using simple technical engineered features.

The LFR is mainly envisioned for electricity and hydrogen production and high-level radioactive material management. The proposed LFR options include a long refueling interval battery ranging from 50 to

150 MWe, a modular system from 300 to 400 MWe, and a large monolithic plant at 1,200 MWe. The LFR battery option (like SSTAR) is a small factory-built turnkey plant operating on a closed fuel cycle with very long refueling interval (15–20 years) cassette core or replaceable reactor module. Its features are designed to meet market opportunities for electricity production on small grids, and for developing countries that may not wish to deploy an indigenous fuel cycle infrastructure to support their nuclear energy systems. Its small size, reduced cost, and full support fuel cycle services can be attractive for these markets. It had the highest evaluations to the GEN-IV goals among the LFR



GEN-IV Reactors. Figure 4
SSTAR-A US lead-cooled fast reactor

options, but also the largest R&D needs and longest development time.

The options in the LFR class may provide a time-phased development path: the nearer-term options focus on electricity production and rely on more easily developed fuel, clad, and coolant combinations and their associated fuel recycle and refabrication technologies. The longer-term option seeks to further exploit the inherently safe properties of lead and raise the coolant outlet temperature sufficiently high to enter markets for hydrogen and process heat, possibly as merchant plants.

The technologies employed are extensions of those currently available from the Russian submarine lead-bismuth alloy-cooled reactors, from the Integral Fast Reactor (IFR [15]) metal alloy fuel recycle and refabrication development, and from the Advanced Liquid Metal Reactor (ALMR [16]) passive safety and modular design approach. Existing ferritic stainless steel and metal alloy fuel, which are already significantly

developed for sodium fast reactors, are adaptable to lead-bismuth-cooled reactors at reactor outlet temperatures of 550°C.

Corrosion of structural materials in lead is one of the main issues for the LFR. Recent experiments confirm that corrosion of steels strongly depends on the operating temperature and dissolved oxygen. Indeed, at relatively low oxygen concentration, the corrosion mechanism changes from surface oxidation to dissolution of the structural steel. Moreover, relationships between oxidation rate, flow velocity, temperature, and stress conditions of the structural material have been observed as well. The compatibility of ferritic and austenitic steels with lead has been extensively studied and it has been demonstrated that generally below 450°C, and with an adequate oxygen activity in the liquid metal, both types of steels build up an oxide layer which behaves as a corrosion barrier. However, above about 500°C, corrosion protection through the oxide barrier appears to fail and is being addressed with various candidate materials. The prospects for extending much above this temperature are not proven at this time.

MSR – Molten Salt Reactor

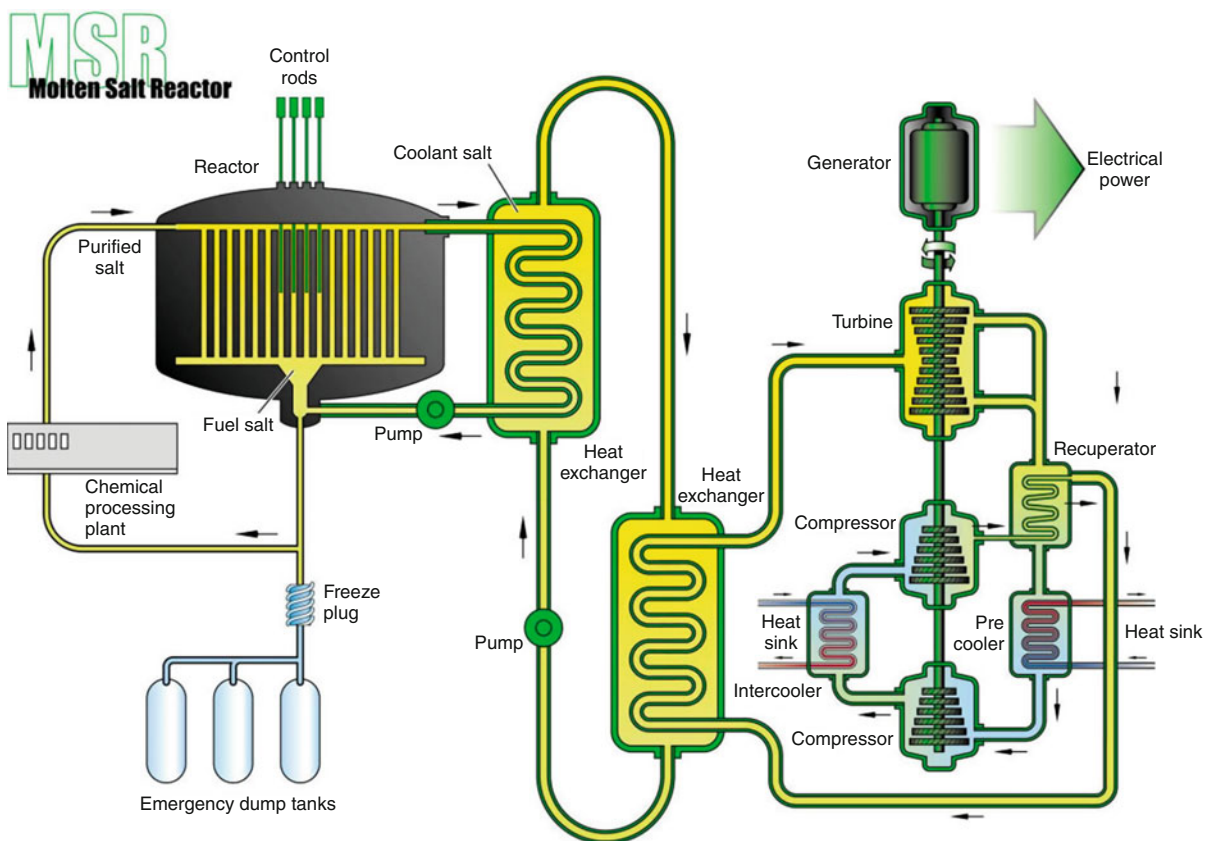
The Molten Salt Reactor uses a molten salt mixture as a primary coolant. Systematic analysis of parameters such as reprocessing time, moderation ratio, core size, and content of heavy nuclei in the salt has resulted in several attractive reactor configurations, in thermal, epithermal, or fast neutron spectrum. The use of a molten salt coolant in a solid-fuel system has been investigated, known as the Advanced High-Temperature Reactor (AHTR [17]), which adapts VHTR fuel form and heat exchanger technology. However, in most MSRs, the fuel is dissolved in the molten salt coolant. Thus, the MSR has unique characteristic compared to other GEN-IV systems: i.e., online refueling and reprocessing are allowed without reactor shutdown because the fuel can move. In addition, the MSR have the following characteristics, which may afford advances: good neutron economy and alternatives for actinide burning or conversion, potential for hydrogen production with high operating temperature, low stresses on the vessel and piping with a very low vapor pressure, enhanced safety by fail-safe drainage,

passive cooling, and a low inventory of volatile fission products, etc. Figure 5 shows the schematic of the MSR concept with dissolved fuel.

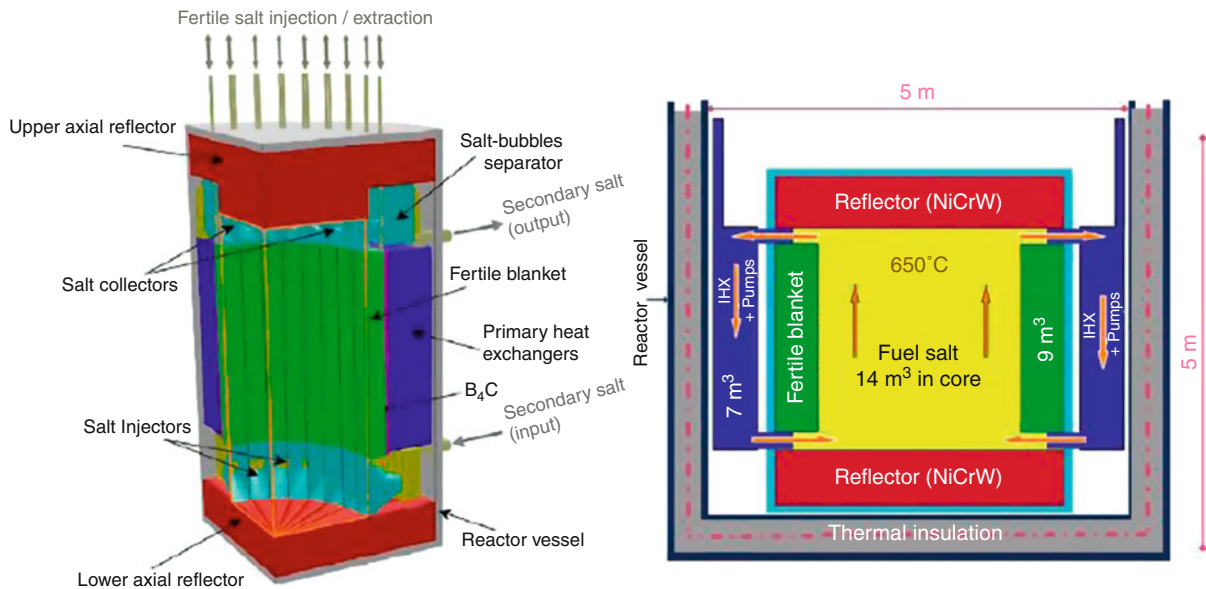
The MSR was first developed in the late 1940s and 1950s for aircraft propulsion. The Aircraft Reactor Experiment (ARE [18]) was a 2.5 MWth nuclear reactor experiment designed to attain a high-power density for use as an engine in a nuclear powered bomber. One experiment used the molten fluoride salt $\text{NaF-ZrF}_4\text{-UF}_4$ (53-41-6 mol%) as fuel, was moderated by beryllium oxide, used liquid sodium as a secondary coolant, and had a peak temperature of 860°C. It operated for a 1,000 h cycle in 1954. The 8 MWth Molten Salt Reactor Experiment (MSRE [19]) was operated from 1965 to 1969 to demonstrate many features, including lithium/beryllium fluoride salt, graphite moderator, stable performance, off-gas systems, and

use of different fuels such as U-233, U-235, and plutonium.

Recently, two MSRs were proposed: Thorium Molten Salt Reactor (TMSR [12]), and FUJI mini-MSR [12]. Figure 6 shows the 1,000 MWe TMSR with graphite moderator. Its operating temperature is 630°C and its thermodynamic efficiency is 40%. The salt used is a binary salt, LiF-(HN)F_4 , with the $(\text{HN})\text{F}_4$ content set to 22%, corresponding to a melting temperature of 565°C. The U-233 enrichment is about 3%. A graphite radial blanket surrounds the core to improve breeding performance. The reprocessing time of the total salt volume is specified to be 6 months, with external storage of the Pa and complete extraction of the fission products and TRU. It is assumed that the U-233 produced in the blanket is also extracted every 6 months. The FUJI mini-MSR is a 100 MWe molten-salt-fueled thorium fuel cycle



GEN-IV Reactors. Figure 5
Molten salt reactor



GEN-IV Reactors. Figure 6
Thorium molten salt reactor

thermal breeder reactor being developed internationally by Japanese, Russian and US consortium. Like all molten salt reactors, the core is chemically inert under low pressures to prevent explosions and toxic releases.

There are four fuel cycle options: (1) maximum breeding ratio (up to 1.07) using a Th and U-233 fuel cycle, (2) denatured Th and U-233 converter with minimum inventory of nuclear material suitable for weapons use, (3) denatured once-through actinide burning (Pu and minor actinides) fuel cycle with minimum chemical processing, and (4) actinide burning with continuous recycling. The fourth option with electricity production is favored for the GEN-IV MSR. Fluoride salts with higher solubility for actinides such as NaF/ZrF₄ are preferred for this option. Salts with lower potential for tritium production would be preferred if hydrogen production was the objective. Lithium and beryllium fluorides would be preferred if high conversion was the objective. On-line processing of the liquid fuel is only required for high conversion to avoid parasitic neutron losses of Pa-233 that decays to U-233 fuel. Off-line fuel salt processing is acceptable for actinide management and hydrogen or electricity generation missions.

The reactor can use U or Th as a fertile fuel dissolved as fluorides in the molten salt. Due to the

thermal or epithermal spectrum of the fluoride MSR, Th achieves the highest conversion factors. However, before sufficient fissile is bred for maintaining the criticality, the MSR requires low-enriched uranium or other fissile materials. The operating temperature ranges from the melting point of eutectic fluorine salts (about 450°C) to below the chemical compatibility temperature of nickel-based alloys (about 800°C).

The R&D will focus on fuel salt cleanup, including pyrochemical separation technologies, extraction of gaseous fission products and noble metals by gas bubbling, tritium speciation and control, and conversion of various waste streams into final waste forms. The research will gradually advance from laboratory scale to larger and more integrated demonstrations. MSR burner and breeder fuel cycles will be evaluated and compared with other nuclear systems. This includes examination of the burning of actinides from other nuclear systems, startup of MSRs on various actinides, avoidance of the generation of most actinides by use of thorium fuel cycles, and alternative breeder reactor fuel cycles.

The MSR also addresses research related to the compatibility of fuel and coolant salts with core and structural materials and challenging MSR subsystem integrity: reactor components and reprocessing unit

regarding mechanical and corrosion resistance. The high temperature, salt reduction-oxidation potential, radiation fluence, and energy spectrum pose a serious challenge for any structural alloy in an MSR. The design of a practical system demands the selection of salt constituents such as LiF, NaF, BeF₂, UF₄, ThF₄, and PuF₃ that are not appreciably reduced by available structural metals and alloys whose component Fe, Ni, and Cr can be in near equilibrium with the salt. Small levels of impurities in the salt may also aggressively corrode the metallics.

Circulating fuel raises challenges within the core such as the loss of delayed neutrons, temperature differences between the salt, reflectors, and moderator, which requires the coupling between neutronics, thermal-hydraulics, salt composition, and properties of the MSR.

SFR – Sodium-cooled Fast Reactor

The Sodium-cooled Fast Reactor features a fast-spectrum reactor and closed fuel-recycle system. Including electricity generation, the primary mission for the SFR could be either enhancement of the uranium resource utilization or high-level radioactive material management, which depends on the SFR designs. Historically, the enhancement of the uranium resource utilization was the primary mission of the SFR by achieving a high breeding ratio, but the mission was recently shifted for consuming transuranics (plutonium and other long-lived radioactive material) in a very low breeding ratio core. The latter has been studied under the Global Nuclear Energy Partnership (GNEP), which was initiated to seek worldwide consensus on enabling expanded use of economical carbon-free nuclear energy to meet growing electricity demand. The GNEP adopted a fully closed nuclear fuel cycle option that enhances energy security while improving proliferation risk management. One of the major goals of the GNEP is to design and demonstrate a SFR for actinide management like the Advanced Burner Reactor (ABR, [20]).

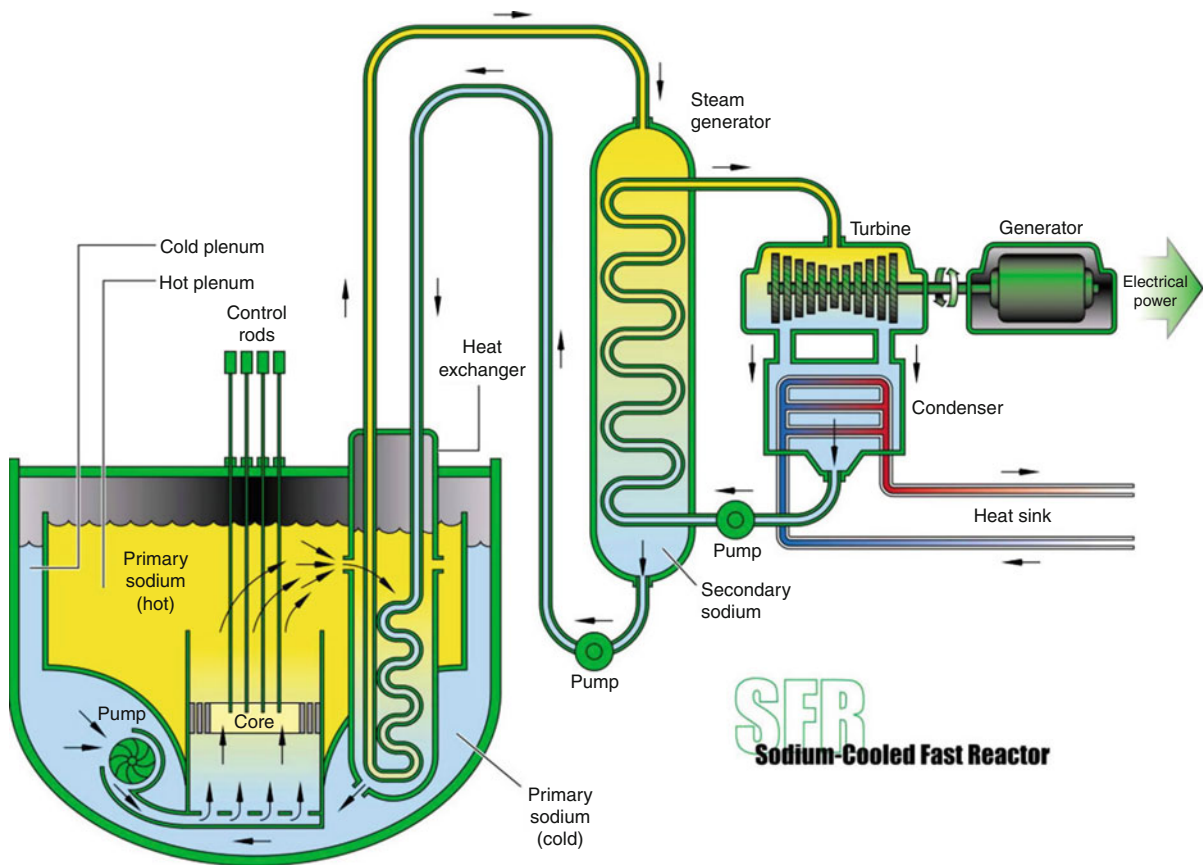
Based on the arrangement of the primary coolant pump and intermediate heat exchanger (IHX), there are two options for the SFR systems: pool type and loop type (see Figs. 7 and 8). The primary pump and IHX are placed inside the reactor vessel in the pool type,

while these two components are located outside reactor vessel by connecting them through pipes. A hybrid option [21] of the pool and loop types has also been proposed.

The experiences on design, construction, and operation provide important input into the design process and have the potential to influence the maturity of the various fast reactor concepts. The greater the number of operating experience years, the greater the opportunity to modify the design based on operating lessons learned. The SFR relies on technologies already developed and demonstrated for sodium-cooled reactors and associated fuel cycles that have successfully been built and operated in worldwide fast reactor programs. Overall, approximately 300 reactor years of operating experience have been logged on SFRs including 200 years on smaller test reactors and 100 years on larger demonstration or prototype reactors. Thus, the technical readiness level, which indicates how soon a system could be deployed, of the SFR is most matured among the six GEN-IV systems.

In the USA, the SFR technology was employed in the 20 MW-electric Experimental Breeder Reactor II (EBR-II [22]) that operated from 1963 to 1994. EBR-II R&D included development and testing of metal fuel and passive safety tests. The 400 MWth Fast Flux Test Facility (FFTF [23]) was completed in 1980 (Fig. 9). The FFTF operated successfully for 10 years with a full core of mixed oxide (MOX) fuel and performed SFR materials, fuels, and component testing. The US SFR development program stalled with cancellation of the Clinch River demonstration reactor in 1983, although US-DOE research for advanced SFR technology continued until 1994. The SFR experience also extends to the commercial sector with the operation of Detroit Edison's FERMI-1 plant from 1963 to 1972.

Significant SFR research and development programs are being conducted in China, France, India, Japan, Russia, and Republic of Korea. The most modern fast reactor construction project was the 280 MWe MONJU (Japan) that was completed in 1990, which will be restarted soon. The construction of 20 MWe Chinese Experimental Fast Reactor (CEFR) and coolant sodium loading was completed in 2009, and the full power operation is expected in 2010. India operates 40 MWth Fast Breeder Test Reactor (FBTR) since 1985 and 500 MWe Prototype Fast Breeder Reactor



GEN-IV Reactors. Figure 7
Pool-type sodium-cooled fast reactor

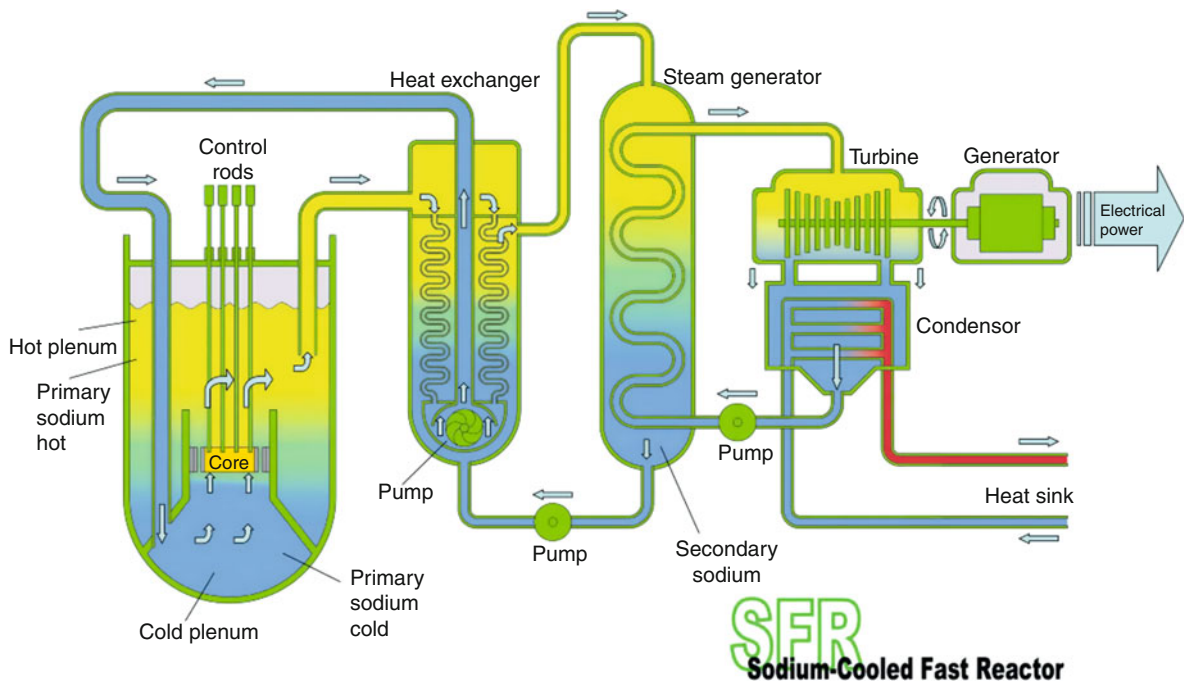
(PFBR) is under construction. The only current fast reactor for electrical generation is the Russian BN-600 that has reliably operated since 1980, and the BN-800 is under construction.

A range of plant size options are available for the SFR, ranging from a battery type systems of a hundred MW-thermal to large monolithic reactors of 3,500 MW-thermal. The sodium coolant outlet temperature is limited by the material properties. Coolant outlet temperatures are typically less than 550°C; however, further increase is considered.

A large margin to coolant boiling is achieved by design, and is an important safety feature of these systems. Another major safety feature is that the primary system operates at essentially atmospheric pressure, pressurized only to the extent needed to move fluid. Sodium reacts chemically with air, and with water, and thus the design must limit the potential for

such reactions and their consequences. To improve safety, a secondary sodium system acts as a buffer between the radioactive sodium in the primary system and the steam or water that is contained in the conventional Rankine-cycle power plant.

Metallic and oxide fuel forms are available for the SFR. The metallic fuel was originally chosen in the early fast reactor programs because of its high density, compatibility with the liquid metal coolant, relative easiness to fabricate, and excellent thermal conductivity. In the late 1960s, before the full potential of metallic fuels were established, the interest worldwide for fast reactor fuel turned toward the oxide fuel, because the achievable burnup is limited by a large irradiation swelling. However, the development and irradiation test of metallic fuels continued though the 1970s and it was discovered that the metallic fuel can achieve a high burnup by allowing room for fuel to swell. In addition,



GEN-IV Reactors. Figure 8
Loop-type sodium-cooled fast reactor



GEN-IV Reactors. Figure 9
Fast Flux Test Facility

the metallic fuel was focused again in the recent fast reactor programs because of its potential passive safety benefits.

The high burnup potential, rich experiences in commercial water-cooled reactors, and the existence

of established industry for manufacturing were the critical factors that motivated interest in oxide fuel for the liquid-metal-cooled fast reactors. However, the low heavy metal density and low thermal conductivity are the principal disadvantages of the oxide fuel. The low density is unfavorable to implement a compact core and increase the breeding ratio or cycle length. The low thermal conductivity leads to high temperature gradient from fuel to coolant. As a result, the oxide fuel stores significant amount of Doppler reactivity in the normal operation condition and it provides the unfavorable positive reactivity feedback during an unprotected severe accident.

Recently, the mixed carbide and nitride fuels have been given attention as the alternative fuels for sodium-cooled fast reactor on the basis of their high density, compatibility with sodium coolant, high melting temperature, and excellent thermal conductivity although they are ceramic fuel like a mixed oxide fuel.

The SFR require a closed fuel cycle to enable their advantageous actinide management and fuel utilization features. There are two primary fuel cycle technology options: an advanced aqueous process and the

pyroprocess [15] which derives from the term, pyrometallurgical process. Both processes have similar objectives: recovery and recycle of more than 99.9% of the actinides, inherently low decontamination factor of the product, making it highly radioactive, and never separating plutonium at any stage for nonproliferation. These fuel cycle technologies are adaptable to thermal spectrum fuels in addition to serving the needs of the SFR. Thus, the reactor technology and the fuel cycle technology are strongly linked.

Due to the flexibility of the conversion ratio depending on the core design options, the SFR can be operated in three distinct fuel cycle roles. A conversion ratio less than 1 (“burner”) can reduce long-lived radioactive waste. A conversion ratio near 1 can increase the uranium utilization without feeding additional enriched uranium. A conversion ratio greater than 1 (“breeder”) affords a net creation of fissile materials. An appropriately designed fast reactor has flexibility to shift between these operating modes; the desired actinide management strategy will depend on a balance of waste management and resource extension considerations.

Regarding economics, the reduction of the plant capital costs is crucial. A number of innovative SFR design features have been proposed: configuration simplifications, improved Operations & Maintenance (O&M) technology, advanced reactor materials, advanced energy conversion systems, fuel handling, etc.

With regard to reactor safety, technology gaps center around two general areas: assurance of passive safety response and techniques for evaluation of bounding events. The advanced SFR designs exploit passive safety measures to increase reliability. The system behavior will vary depending on system size, design features, and fuel type. R&D for passive safety will investigate phenomena such as axial fuel expansion and radial core expansion, and design features such as self-actuated shutdown systems and passive decay heat removal systems. The ability to measure and verify these passive features must be demonstrated. Associated R&D will be required to identify bounding events for specific designs and investigate the fundamental phenomena to mitigate severe accidents.

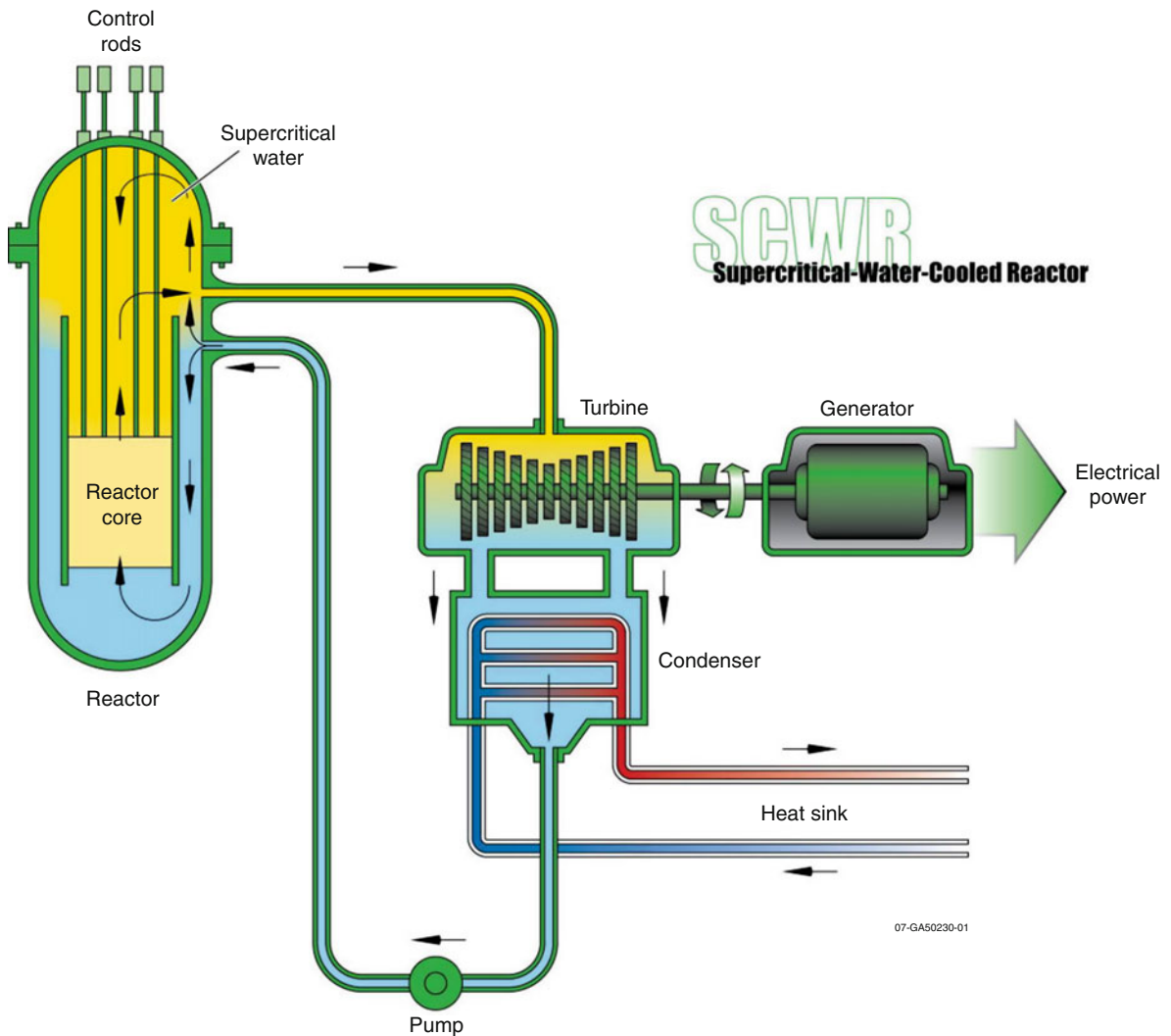
Finally, the development of SFR technology provides the opportunity to design modern safeguards directly into the planning and building of new nuclear

energy systems and fuel cycle facilities. Incorporating safeguards into the design phase for new facilities will facilitate nuclear inspections conducted by the International Atomic Energy Agency (IAEA). The goal of this oversight is to always have an accurate grasp of the current inventory through the utilization of advanced technologies to verify the characteristics of the security system (accountancy, detection, and promptness) and the physical protection characteristics (physical protection measures, the monitoring level, and security measures) and for ensuring robust design to guarantee these characteristics. It is also necessary to maintain transparency and openness in terms of information to more effectively and efficiently monitor and verify nuclear material inventories.

SCWR – Supercritical Water-cooled Reactor

The Supercritical Water-cooled Reactor is a water-cooled reactor like Light Water Reactor (LWR) operated commercially, but the SCWR is operated above the thermodynamic critical point of water (374°C, 22.1 MPa). Figure 10 shows the SCWR system.

The specific heat increases drastically and the water density decreases without boiling of water around the thermodynamic critical point. As a result, the SCWR has unique features that may offer advantages compared to state-of-the-art PWRs: Higher plant thermal efficiency compared to LWRs due to the higher operating temperature. Low density of water without boiling allows the direct cycle like Boiling Water Reactor (BWR), but steam dryers, steam separators, recirculation pumps, and steam generators are not necessary, and as a result, the SCWR can be a simpler plant with fewer major components. Lower-coolant mass flow rate per unit core thermal power results from the high heat capacity of the supercritical water. This offers a reduction in the size of the reactor coolant pumps, piping, and associated equipment, and a reduction in the pumping power. Lower-coolant mass inventory results from the once-through coolant path in the reactor vessel and the lower-coolant density. This opens the possibility of smaller containment buildings. No boiling crisis (i.e., departure from nucleate boiling or dry out) exists due to the lack of a second phase in the reactor, thereby avoiding discontinuous heat transfer regimes within the core during normal operation.



GEN-IV Reactors. Figure 10
Supercritical water-cooled reactor

The SCWR systems may have a thermal [24], fast [25], or mixed-neutron spectrum [26] depending on the core design. The Japanese supercritical light water reactor (SCLWR) with a thermal spectrum has been the subject of the most development work in the last 10–15 years. The SCLWR reactor vessel is similar in design to a PWR vessel (although the primary coolant system is a direct-cycle, BWR-type system). High-pressure (25.0 MPa) coolant enters the vessel at 280°C. The inlet flow splits, partly to a downcomer and partly to a plenum at the top of the core to flow down through the core in

special water rods. This strategy provides moderation in the core. The coolant is heated to about 510°C and delivered to a power conversion cycle, which blends LWR and supercritical fossil plant technology; high-, intermediate-, and low-pressure turbines are employed with two reheat cycles.

The SCWR can also be designed to operate as a fast reactor. The difference between thermal and fast versions is primarily the amount of moderator material in the SCWR core. The fast spectrum reactors use no additional moderator material, while the thermal

spectrum reactors need additional moderator material in the core. The mixed-spectrum SCWR was proposed not only to achieve all advantages of SCWR but also the actinide management. The core uses two coolant flow paths: outer zone with high density water and inner zone with low density water (see Fig. 11). Thus, the inner zone features fast neutron spectrum, while the outer zone features thermal spectrum. By recycling TRU in the fast zone, the mixed-spectrum SCWR is capable of keeping all TRU in the reactor.

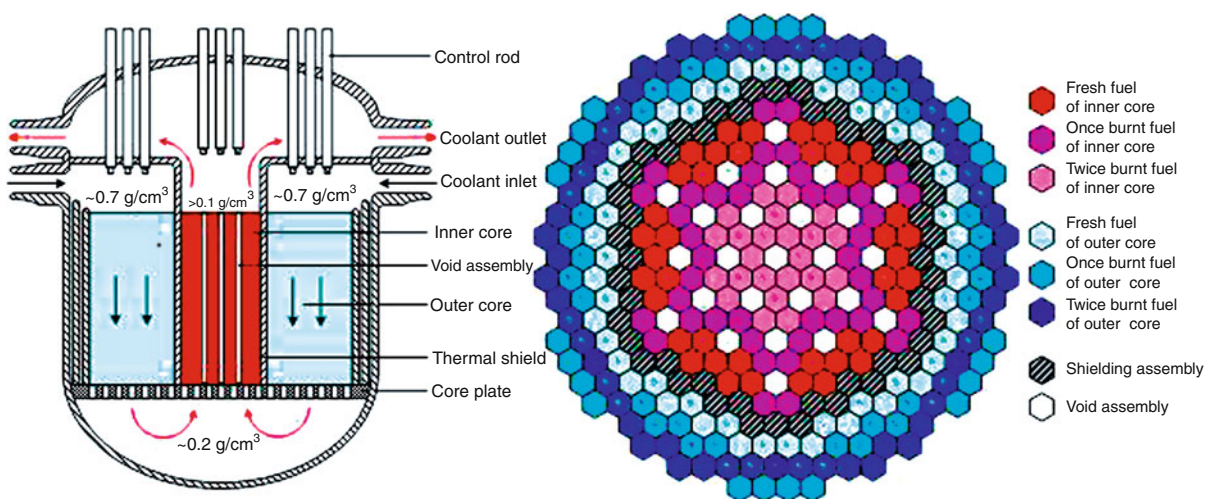
Much of the technology base for the SCWR can be found in the existing LWRs and in commercial supercritical-water-cooled fossil-fired power plants. However, there are some relatively immature areas. There have been no prototype SCWRs built and tested. For the reactor primary system, there has been very little in-pile research done on potential SCWR materials or designs, although some SCWR in-pile research has been done for defense programs in Russia and the United States. Limited design analysis has been underway over the last decade in Japan, Canada, and Russia. For the balance of plant, there has been development of turbine generators, piping, and other equipment extensively used in supercritical-water-cooled fossil-fired power plants.

The ability to use proven uranium oxide fuel greatly simplifies the application of fuel and fuel cycle

technology to the SCWR. However, the supercritical water is known to challenge the corrosion/erosion performance of current cladding technology, and R&D is focused on advanced cladding materials.

There are several unique components needed for the SCWR, including the reactor pressure vessel or pressure tubes and its internal structural components, moderator channels, control rods and drives, the condenser and high-pressure pumps, valves, and seals. The reactor pressure boundary must operate above the high pressure (22.1 MPa) of supercritical water. This may be addressed with thicker sections, and thermal stresses can be avoided with a thermal sleeve for the outlet nozzle.

Zirconium-based alloys, common in water-cooled reactors, may not be a viable material without thermal and/or corrosion-resistant barriers. Based on available data for other alloy classes, there is no single alloy that has received enough study to unequivocally ensure its performance in an SCWR. Another key need of this system will be an enhanced understanding of the chemistry of supercritical water. Water above its critical point is accompanied by dramatic changes in chemical properties. Its behavior and degradation of materials is further accelerated by in-core radiolysis, which preliminary studies suggest is markedly different than what would have been predicted by simplistic extrapolations from conventional reactors.



GEN-IV Reactors. Figure 11

Core layout of mixed-spectrum supercritical water reactor

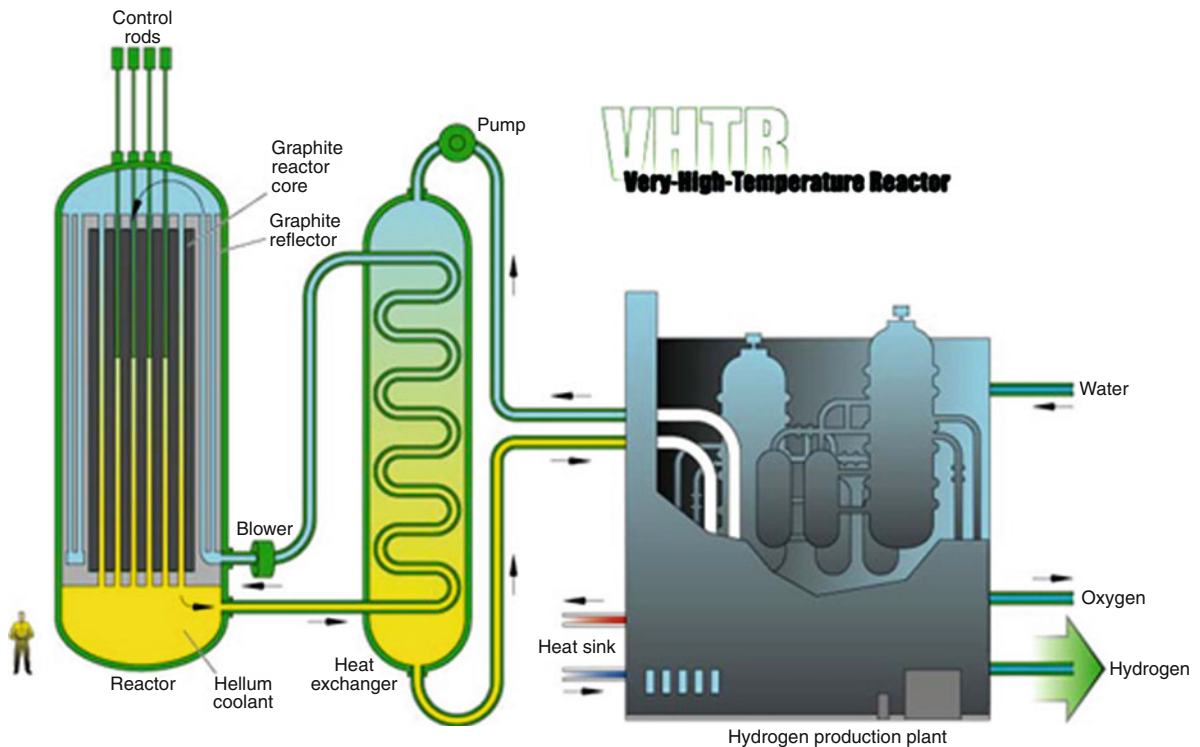
The approach to development of materials and components will build on evaluation of candidate materials with regard to corrosion and stress corrosion cracking, strength, embrittlement and creep resistance, and dimensional and microstructural stability; the potential for water chemistry control to minimize impacts as well as rates of deposition on fuel cladding and turbine blades; and measurement of performance data in an in-pile loop. All of these are critical to establishing viability of the SCWR.

The SCWR leads the way among GEN-IV systems in the development of advanced materials for water coolant. In fact, the diffusion of this technology into current generation light and heavy water reactors seems assured. However, much remains to be done: the thermal-hydraulic performance during normal and off-normal operation, as well as postulated accidents, needs to be addressed both with advances in the design and safety approach as well as the analysis tools. Issues to be addressed include the basic thermal-hydraulic

phenomenon of heat transfer and fluid flow of super-critical water in various geometries, critical flow measurements, the strong coupling of neutronic and thermal-hydraulic behavior, leading to concerns about flow stability and transient behavior, validation of computer codes that reflect these phenomena, and definition of the safety and licensing approach as distinct from current water reactors, including the spectrum of postulated accidents, flow instability, etc.

VHTR – Very-high-temperature Reactor

The Very-high-temperature Reactor is a graphite-moderated, helium-cooled reactor like GT-MHR and PBMR capable of generating electricity, but the coolant output temperature is significantly increased up to 1,000°C. In Fig 12, the schematic of the VHTR is depicted. The higher temperatures of this reactor open the door for industrial heat processing opportunities, in particular, for hydrogen production.



02-GA50657-01

GEN-IV Reactors. Figure 12
Very-high-temperature reactor

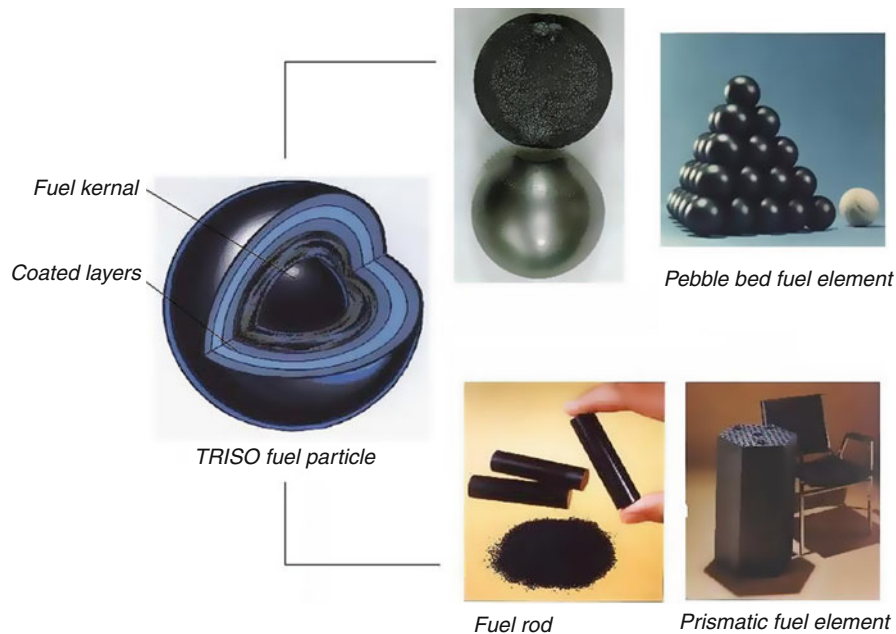
The annual US demand for hydrogen is over 12 million tons, and expected to grow to over 30 million tons by 2030. Industry uses hydrogen for fossil fuel refining, treating metals, and food processing. Hydrogen is currently produced primarily from steam methane reforming using fossil fuel as a heat source. Hydrogen can also be produced by various processes using a high-temperature gas-cooled reactor as the primary energy source.

Use of nuclear energy as the heat source of a large-scale hydrogen production operation would result in substantially lower carbon emissions over a natural gas-fired steam methane reforming operation. A 600 MWth VHTR dedicated to hydrogen production can yield over 2 million normal cubic meters per day. The VHTR can also generate electricity with high efficiency, over 50% at 1,000°C.

The VHTR has been evolved from gas-cooled reactor experiences and extensive international databases that can support its development. The basic technology for the VHTR has been well established in former gas-cooled reactors, such as DRAGON, Peach Bottom, AVR, THTR, and Fort St Vrain, and is being advanced

in concepts such as the GT-MHR and PBMR. The ongoing 30-MWth HTTR project in Japan is intended to demonstrate the feasibility of reaching outlet temperatures up to 950°C coupled to a heat utilization process, and the HTR-10 in China will demonstrate electricity generation at a power level of 10 MWth. The former projects in Germany and Japan provide data relevant to the VHTR development.

The VHTR core uses TRISO particles to form a pebble bed or prismatic fuel element (see Fig. 13). The TRISO particle, which has a small diameter of less than 1.0 mm, has a fuel kernel in the form of uranium oxide. The enrichment of the uranium is dependent on the core design purposes. The kernel is subsequently coated with a porous carbon layer (to hold fission gases), a dense pyrolytic carbon layer, a silicon carbide layer, and finally another pyrolytic carbon layer. The coatings surrounding the kernel of TRISO particles produce a very robust fuel form by acting as the containment boundary for the radioactive material. These coatings work in much the same way as the massive reinforced concrete structure surrounding the light water reactors currently in service.



GEN-IV Reactors. Figure 13
VHTR fuel elements

The reactor core type of the VHTR can be a prismatic block core such as GT-MHR and Japanese HTTR, or a pebble-bed core such as PBMR and Chinese HTR-10. Despite of the alternate fuel element designs (pebble bed versus prismatic), the two baselines have many technologies in common that allow for a unified R&D approach. The well-known TRISO particle fuel with a UO₂ kernel and SiC/PyC coating may be used in either, or it may be enhanced with a different fuel kernel form such as UCO or an advanced ZrC coating through additional research. For electricity generation, the helium gas turbine system can be directly set in the primary coolant loop, which is called a direct cycle. For nuclear heat applications such as process heat for refineries, petro-chemistry, metallurgy, and hydrogen production, the heat application process is generally coupled with the reactor through an intermediate heat exchanger (IHX), which is called an indirect cycle.

The fuel cycle will initially be a once-through fuel cycle specified for high burnup (15–20 atom-%) using low enriched uranium. The operation with a closed fuel cycle will be assessed and solutions to better manage the fuel cycle back end will be developed. The possible use of TRU as a fuel will be studied conceptually for actinide management [27].

The primary emphasis in fuel development is on its performance at high burnup, power density, and temperature. The R&D broadly addresses its manufacture and characterization, irradiation performance, and accident behavior. Irradiation tests will provide data on coated particle fuel and fuel element performance

under irradiation as necessary to support fabrication process development, to qualify the fuel design, and to support development and validation of models and computer codes on fission product transport. They will also provide irradiated fuel and materials samples for postirradiation and safety testing. The performance expected for the fuel must be verified for all normal, transient, or accident conditions as well as certain severe accident conditions (beyond design basis). A key claim of the fuel is its ability to retain fission products in the fuel particles under a range of postulated accidents with temperatures up to 1,600°C.

Future Directions

The objective for Generation IV nuclear energy systems is to have them available for wide-scale deployment before the year 2030. The anticipated deployment dates for the six GEN-IV systems are provided in Table 3 in terms of R&D phases. The deployment dates of the SFR and VHTR are expected to be earlier than other GEN-IV systems because of their matured technical readiness level.

In the viability R&D phase, the feasibility of key technologies of the GEN-IV systems will be examined. The performance R&D activities undertake the development of performance data and optimization of the system. Assuming the successful completion of viability and performance R&D, the demonstration R&D phase activities involve the licensing, construction, and operation of a prototype or demonstration system in partnership with industry and perhaps other countries.

GEN-IV Reactors. Table 3 Anticipated deployment of GEN-IV systems

System	Deployment timelines		
	Viability phase	Performance phase	Demonstration phase
GFR	2012	2020	2025
LFR	2014	2020	2025
MSR	2013	2020	2025
SFR	2006	2015	2020
SCWR	2014	2020	2025
VHTR	2010	2015	2020

Thus, the detailed design and licensing of the system will be performed during the demonstration phase. The R&D projects and milestones anticipated in each phase were defined in GEN-IV roadmap project [2].

Bibliography

Primary Literature

1. Generation IV International Forum. <http://www.gen-4.org>
2. US-DOE Nuclear Energy Research Advisory Committee and GIF (2002) A technical roadmap for generation IV nuclear energy systems: ten nations preparing today for tomorrow's energy needs. Generation IV International Forum. <http://www.gen-4.org>
3. Potter PC (1996) Gas turbine-modular helium reactor (GT-MHR) conceptual design description report. General Atomics GA-910720
4. Koster A, Matzer HD, Nichols DR (2003) PBMR design for the future. Nucl Eng Des 222:231–245
5. Simon R (2005) The primary circuit of the DRAGON high temperature reactor experiment. 18th International conference of structural mechanics in reactor technology (SMIRT 18), Beijing
6. Moormann R (2008) A safety re-evaluation of the AVR pebble bed reactor operation and its consequences for future HTR concepts. Institute for Energy Research, Switzerland
7. Wachholz W (1988) The present state of the HTR concept based on experience gained from AVR and THTR. International Working Group on Gas-cooled Reactors International Atomic Energy Agency IWGGCR-19, Vienna
8. Brown JR et al (1987) Physics testing at Fort St. Vrain: a review. Nucl Sci Eng 97:104
9. Yamashita K et al (1996) Nuclear design of the high-temperature engineering test reactor (HTTR). Nucl Sci Eng 122:212
10. Seker V, Colak U (2003) HTR-10 full core first criticality analysis with MCNP. Nucl Eng Des 222:263–270
11. Stainsby R, Peers K, Mitchell C, Poette C, Mikityuk K, Somers J (2009) Gas cooled fast reactor research and development in the European Union. Sci Technol Nucl Installations 2009:1–7
12. IAEA (2007) Status of small reactor designs without on-site refueling. International Atomic Energy Agency IAEA-TECDOC-1536, Vienna
13. Smith C, Halsey WG, Brown NW, Sienicki JJ, Moiseyev A, Wade DC (2008) SSTAR: the US lead-cooled fast reactor (LFR). J Nucl Mater 376:255–259
14. Cinotti L, Smith CF, Sienicki JJ, Abderrahim HA, Benamati G, Locaelli G, Monti S, Wider H, Stuwe D, Orden A (2007) The potential of LFR and ELSY project. 2007 International congress on advances in nuclear power plants, Nice
15. Hannum WH (1997) The technology of Integral Fast Reactor and its associated fuel cycle. Prog Nucl Energy 31: 1–217
16. Boardman CE (2000) A description of the S-PRISM plant. 8th International conference on nuclear engineering, Baltimore
17. Ingersoll T et al (2004) Status of pre-conceptual design of the Advanced High-Temperature Reactor (AHTR). Oak Ridge National Laboratory, Oak Ridge, ORNL/TM-2004/104
18. Bettis ES, Cottrell WB, Mann ER, Meem JL, Whitman GD (1957) The aircraft reactor experiment: operation. Nucl Sci Eng 2:841–853
19. Haubenreich PN, Engel JR (1970) Experience with the molten-salt reactor experiment. Nucl Appl Technol 8:118–136
20. Kim TK, Yang WS, Grandy C, Hill RN (2009) Core design studies for a 1000 MWt advanced burner reactor. Ann Nucl Energy 36:331–336
21. Zhao H, Zhang H (2007) An innovative hybrid loop-pool design for sodium cooled fast reactor. Idaho National Laboratory INL/CON-07-12657
22. Koch L (2000) Experimental breeder reactor-II. Argonne National Laboratory, Chicago
23. Lash T (1997) Fast flux test facility (FFTF) briefing book 1: summary. Pacific Northwest National Laboratory PNNL-11778
24. Oka Y (2000) Design concept of once-through cycle supercritical pressurized light water reactors. The first International symposium on supercritical water-cooled reactors, SCR-2000, Tokyo
25. Macdonald P (2000) Feasibility study of supercritical light water cooled fast reactor for actinide burning and electric power production. Idaho National Laboratory INEEL/EXT-02-01330
26. Kim TK, Wilson PH, Hu P, Jain R (2004) Feasibility and configuration of a mixed spectrum supercritical water reactor. International topical meeting on the Physics of Fuel Cycles and Advanced Nuclear Systems (PHYSOR-2004), Chicago
27. Kim TK, Taiwo TA, Yang WS, Hill RN, Assessment of deep burnup concept based on graphite moderated gas-cooled thermal reactor. International topical meeting on the Physics of Fuel Cycles and Advanced Nuclear Systems (PHYSOR-2006), Vancouver

Book and Reviews

- A Strategy for nuclear energy research and development. Electric Power Research Institute (EPRI)
- Bouchard JB (2008) Generation IV advanced nuclear energy systems. Nucl Plant J 26
- Kim WJ et al (2006) Supercritical carbon dioxide Brayton power conversion cycle design for optimized battery-type integral reactor system. International congress on advances in nuclear power plants, Reno
- MacDonald PE et al (2003) NGNP preliminary point design – results of initial neutronics and thermal-hydraulic assessment. INEEL/EXT-03-00870 Rev. 1. Idaho National Engineering and Environmental Laboratory, September 2003

- Sienicki JJ et al (2006) Status report on small secure transportable autonomous reactor (SSTAR)/lead-cooled fast reactor (LFR) and supporting research and development. Argonne National Laboratory ANL-GenIV-089
- Schultz RR (2008) Next generation nuclear plant methods research and development technical program plan. Idaho National Laboratory INL/EXT-06-11804
- The US Generation IV Implementation Strategy (2003) US Department of Energy office of Nuclear Energy, Science and Technology
- The US Generation IV Fast Reactor Strategy (2006) US Department of Energy office of Nuclear Energy

Genotype by Environment Interaction and Adaptation

IGNACIO ROMAGOSA¹, GISELA BORRÀS-GELONCH¹,

GUSTAVO SLAFER¹, FRED VAN EEUWIJK²

¹Department of Crop and Forest Sciences, University of Lleida, Lleida, Spain

²Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands

Article Outline

Glossary
 Definition
 Introduction
 Breeding Implications
 Traits Determining Adaptation
 Statistical Approaches for GE Characterization
 Future Directions
 Bibliography

Glossary

GE *Genotype by environment interaction* is differential genotypic expression across environments that may cause that a genotype selected among the best in one location to perform poorly in another. GE weakens association between phenotype and genotype, reducing genetic progress in breeding programs. In statistical terms, GE describes a situation in which the simultaneous effect of two classification variables (genotype and environment) on a continuous dependent third one, such as yield, does not follow an additive model.

MET A *multi-environment trial* is a series of trials sampling the target environmental range in which a particular set of genotypes is evaluated.

QTL A *quantitative trait locus* is a region in the genome associated with a particular quantitative phenotypic trait, such as crop yield, resource-use-efficiency, phenology, or height. QTL analysis is a statistical method that links phenotypic data (specific trait measurements on a series of individuals) and genotypic data (usually in the form of molecular markers taken on the same individual) in order to explain the genetic basis of complex traits. QTL number and the variation they explain on the phenotypic trait give clues about the genetic control of that trait, for example, if plant height is controlled by many genes of small effect, or by a few genes of large effect.

QTLxGE QTL by environment interaction is differential QTL effect across environments that may cause that a favorable QTL in one environment may become irrelevant, or even unfavorable, in another.

Specific and wide adaptation A genotype is considered stable if it yields well relative to the productive potential of the environments in which is grown. If such concept of stability is shown for a wide agroecological array of environments, a genotype is considered to have general, wide, or broad adaptation. If stability is confined to a limited range, a genotype is said to have specific or narrow adaptation.

Definition

One of the first decisions farmers have to take is the selection of the variety to be grown in their fields based on expectation of economic returns, generally, in the form of the highest attainable yield. This is a critical choice that strongly determines the sustainability of the agricultural system. However, this is by no means trivial as it is very hard to identify the “best” variety across a diverse set of environments subjected to complex biotic and abiotic factors and interactions generally causing significant changes in varietal rank. Therefore, a major objective in plant breeding programs is to determine the potential adaptation of advanced breeding lines across a range of agroecological conditions. William S. Gosset (who signed as “Student [1]”

in a landmark publication introducing the *t* distribution) wrote at the onset of modern breeding that the ultimate purpose of field experimentation was to determine what varieties pay farmers best. He thought that the design of experiments should aim, not only at determining the average yield, but also at identifying varieties whose yield, being within those of high average value, were relatively less responsive to variation in soil and climate.

Breeding programs normally aim to release cultivars to be successfully grown over a rather large cropping area, varying in soil quality attributes and in average climate, and across several growing seasons, with interannual variations in climatic conditions. The target environment is defined as the set of soil \times climatic conditions in which the released cultivars will be grown and to which the cultivars must be adapted. Therefore, a key step in applied plant breeding is the identification of advanced genotypes broadly or narrowly adapted across a wide range of target environments. Breeders focus in the first segregating generations on direct phenotypic selection of highly heritable traits, such as plant architecture and phenology to concentrate in later stages on complex quantitative traits like yield and end-use quality. Marker-assisted selection aims at complementing this phenotypic selection with direct marker screening for, mostly, oligogenic-controlled traits. The traditional approach to estimate the genotypic value in the context of breeding, varietal registration, and recommendation is deployment of extensive field evaluation schemes in a series of sites in which the assessed genotypes could be potentially grown. These collections of trials are generally denominated multi-environment trials (METs) in which a set of genotypes is evaluated in a series of trials that sample the target environmental range. Data from METs are typically summarized in the form of genotype by environment tables of means. Simple inspection of such tables of means will often reveal the presence of genotype by environment interaction (GE) or differences in performance of genotypes that are trial dependent. They also allow for the identification of those genotypes that are partially or generally adapted to the environmental range, showing specific or narrow versus general or wide adaptation, respectively.

The traditional outcome of METs is the identification of “which” cultivar and “where” has performed

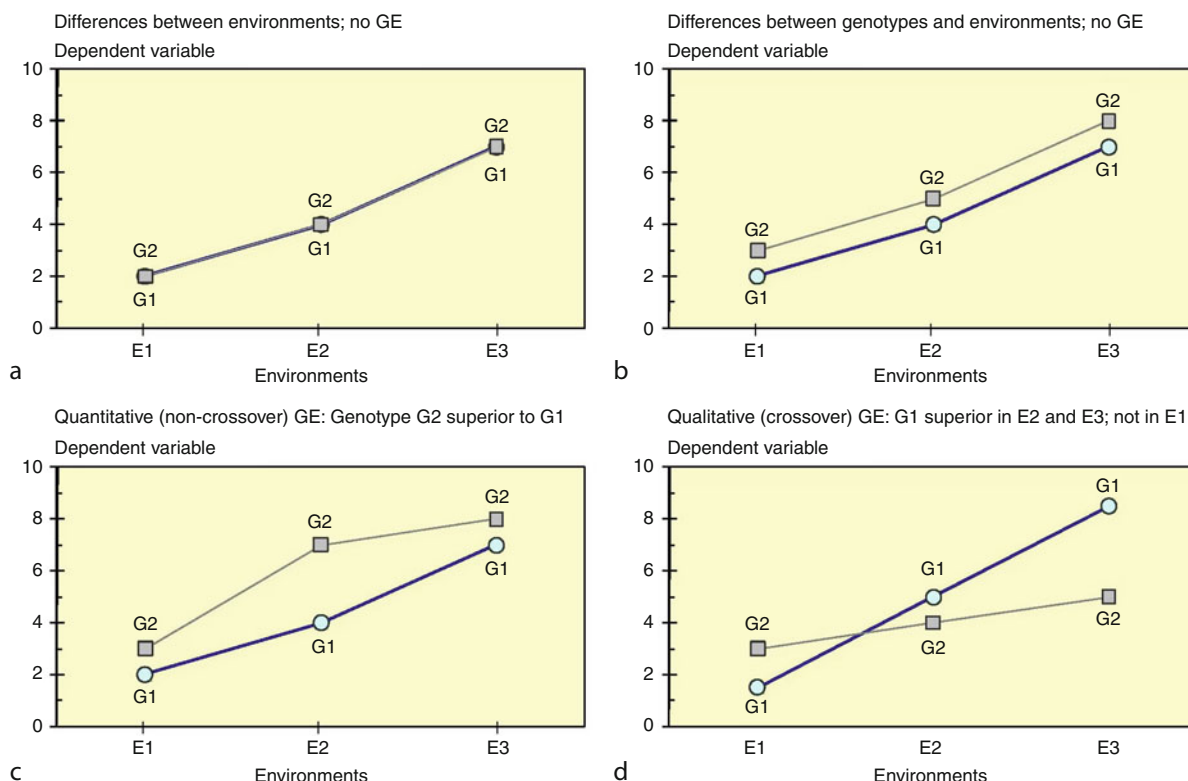
well. These studies are empirical, based on simple statistical characterizations of genotypic responses across environments and do not provide any physiological insight into the basis of the genotypic response to environmental changes. However, as one wants to move forward toward a *predictive* breeding scenario, the challenge beyond “which” and “where” is “why” narrow or wide adaptation happens, in terms of a thorough understanding of both the environment, the physiological behavior of the different cultivars and, eventually, of the genes responsible for adaptation. Identifying the “why” is not only a matter of satisfying curiosity: It would potentially allow more precise breeding through the direct manipulation of the genes responsible for the different adaptation patterns.

Introduction

Statistical analyses that detect and describe GE have been comprehensively reviewed [2–14]. Means across environments in METs are only adequate estimates of varietal performance in the absence of GE. When GE is significant, average values across environments may hide subsets of environments where genotypes differ markedly in relative performance.

As for any other statistical two-factor model, there are different types of interactions which originate from departure from additivity. In Fig. 1, the average for each of two genotypes, G1 and G2, for the dependent variable of interest, for example yield, is shown for three environments. Figure 1a represents the situation in which differences were detected only between environments. Figure 1b shows an additive model in which differences for both main effects, genotypes and environments, were observed but no GE. Figure 1c shows a quantitative or non-crossover interaction; in this scenario, genotypes with superior means can be recommended for all environments. In plant breeding, the most important GE is of the crossover or qualitative type (Fig. 1d), which implies changes in the rankings of genotypes across environments). In this case, variety G2 may be recommended for environment E1 but not for E2 and E3.

When there are genotypic differences among the varieties tested and the target environments include different soils and variable climate, MET analyses more often than not detect crossover GE (only MET with limited



Genotype by Environment Interaction and Adaptation. Figure 1

Performance of two hypothetical genotypes in three environments showing: (a) Only environmental differences; (b) No GE; (c) Quantitative, non-crossover, GE; (d) Qualitative, crossover, GE

genotypic and/or narrow environmental diversity might reveal negligible or nonsignificant interaction). Thus, identification of superior genotypes is complicated by qualitative GE and largely depends on extensive field testing conducted over years at different locations. Therefore, there is a strong need to deploy powerful statistical models for MET data taking into account GE and their breeding implications.

Crossover interactions represent a double-edged sword [10]. Whereas they make breeding, testing, selection, and varietal recommendation more difficult, if the underlying ecophysiological grounds of GE are known, identification of genotypes better adapted to certain specific niche conditions, allowing for increased genetic gains, is possible. If the traits conferring adaptation to these specific environments and/or the genes that control them are revealed, direct implementation in breeding may be feasible either by choosing parents

for a new cross possessing the adaptive attributes or by directly selecting for the presence of such attributes in the progenies (through direct measurement of the attributes or through genotypic selection, see below).

METs are often carried out over a number of sites and years that are considered to be representative of the target environments. Standard analyses of variance partition the GE term into genotype by locations (GL), genotype by years (GY), and genotype by locations by years (GLY) interactions. The relative size of these terms allow for a statistical assessment of the spatial and temporal components of adaptation. If GL dominates over the other components, then specific adaptation is exploitable by identifying subsets of homogeneous locations for variety release and recommendation. Where GY and GLY terms dominate, as most often happens, no simplification involving spatial subdivision of growing regions is possible. In this

context, specific recommendations may be only possible after counting with robust models trustworthily predicting the main climatic conditions of the growing season in advance to sowing.

Recent efforts have searched for the genetic factors underlying GE and, thus, to describe adaptation patterns. Quantitative trait loci (QTLs) responsible for individual complex traits (see, e.g., [15]), such as yield and adaptation have been reported in several populations for most crop species. QTL related to adaptation show different effects in different environments. The magnitude of individual QTL effects (expressed as the amount of GE variation explained by a particular QTL) varied among populations and across environments. Therefore, implementation of marker-assisted selection strategies for these QTLs in applied breeding programs remains a challenge. Modern GE studies have introduced external environmental, physiological, and/or genetic information to develop statistical models whose parameters relate better to physiological knowledge [16, 17], and therefore offer better possibilities for implementation of QTL selection methodologies in breeding programs.

Breeding Implications

Historically most of the genetic progress in the last decades at the global level, particularly in cereals, has been attained through increases of yield potential and disease resistance. Genetic gains in yield under non-limited growing conditions, i.e., improving yield potential, have often brought about parallel gains in yield under a wide range of more realistic, largely stressful, growing conditions [18–20]; because physiological traits behind improved yield potential may often be constitutive and provide yield advantage over a range of conditions [21]. Thus, improving simultaneously for yield potential (which is directly linked to both attainable and on-farm yields; [22]) and for disease resistance has conferred not only clear progress under high-yielding conditions but also wide adaptation.

Thus, it is critical to further improve yield potential [23]. Lessons from the past allow to optimistically trusting that relatively simple traits might be found that affect yield potential and wide adaptation simultaneously (e.g., [24]). For instance, the incorporation

of simple key traits such as reduced height might have such a great impact that may be the basis of a Green Revolution due to its capacity of increasing yield both under potential and most non-potential conditions. Genetically reducing the capacity of the stems to grow through introgression of semidwarfing genes determined firstly an increased partitioning of biomass accumulated during stem elongation to the growing spikes [25, 26]; then the additional availability of resources in the growing spikes allowed floret development to proceed normally in more floret primordia consequently increasing the number of grains [27] and therefore parallel improvements in yield, as cereals are most frequently sink-limited during grain filling even under nonoptimal environments [28, 29]. However, as further reducing height would not keep improving yields [30], it is critical identifying alternative traits that being rather simple were still putatively related to yield across a wide range of conditions. Difficulties in identifying such traits is reflected in the fact that despite continuous breeding efforts in the last decade, current genetic progress in yield potential fall short of both those attained before (see [31] and references therein) and that required to match expected increases in demand [23]. Future improvements in yield potential would largely depend upon the identification of alternative traits that being relatively simple are putatively related to yield in a wide range of conditions representing the target environments of the breeding program. In this context, a thorough examination of GE will be critical both for identifying traits in a top-down approach dissecting yield into physiologically sound traits across conditions representing the target environments, and for determining the stability of the relationship between the identified trait(s) and yield.

In an even more general context, GE has important implications in applied breeding programs [5]. Based on the magnitude and nature of GE, breeders have to decide whether to aim for wide or for specific adaptation. This decision determines the choice of locations for selection, the allocation of limited resources in advanced line testing, and the assessment of the potential trade-off between empirical, molecular, and physiological screening of parents and advanced lines. Related to wide adaptation is the question of breeding sites: Can selection under optimum high-input

environments identify genotypes adapted to more stressed environments? Salvatore Ceccarelli and Stefania Grando at ICARDA have produced a significant number of contributions on the issue of wide versus specific adaptation in barley (see [32] and their own references therein for a review). They have strongly advocated the exploitation of specific adaptation for optimum use of resources particularly in marginal environments, arguing that selection for high yield potential has not increased yield under low-input conditions. However, success of the CIMMYT wheat program aiming at wide adaptation is based on a completely different approach. Rather than focusing on any specific environmental conditions, continuous selection cycles, referred to as shuttle breeding, are carried out in alternative and extremely diverse high yield potential environments differing in altitude, latitude, photoperiod, temperature, rainfall, soil type, and disease spectrum. As a result, CIMMYT wheat genotypes have shown high yield potential and wide adaptation across large geographical regions, perhaps with the exception of very marginal; in fact, poor adaptation of CIMMYT genotypes to specific environments often reflected susceptibility to specific plant diseases.

Field experimentation aims at covering a representative sample of environmental variation. However, the need for adequate resource allocation raises the question of whether multilocation testing in a limited number of years can adequately sample the array of environmental conditions where a variety can be grown. If the MET analysis of variance identifies GY as the most significant term, testing for many crop cycles should be preferred. However, this is not suitable given the increasing pressure to develop new cultivars. Therefore, breeders often substitute temporal for spatial environmental variation, assuming that GL is similar in nature to GY and that GLY is absent. Resource allocation for varietal experimentation schemes depends on the relative magnitude of the variance components for the genotype and GE interaction terms. Given the small number of years available for testing, and the frequently dominant effect of GY and GLY interactions, there is little point in a very extensive series of trials in a given year with a high proportion of genotypes retained throughout. Integrated mixed model analyses for the selected genotypes across the breeding stages can counterweigh for the limited number of years in the later stages of field testing.

A series of papers have suggested the use of reference and probe genotypes to characterize environmental variation and assess GE repeatability [33]. By defining a common reference set of genotypes consistently grown across locations and years, a breeder could define a long-term target environment and weight results from each location in a given year in accordance with its across-year representativeness. Probe genotypes with differential response to known biotic and abiotic conditions could also be used to characterize environments. However, practical application of these two principles is not common. Genetic gains for unidentified biotic and abiotic stresses by direct selection on extensive MET are possible. A more sound approach could be the growing of genotypes in a few key environments with well-characterized levels of the target stress. Manipulation of the breeding environment and selection of key parents for crossing should result in improved genetic gains. However, this second approach requires a clear understanding of the major stress as well as the facilities to reproduce it.

A germplasm strategy is also needed for breeding for wide and specific adaptation. For most crops, there is an important gap between elite and unimproved gene pools as most breeders focus on germplasm reflecting decades of intensive crossing, selection, and recombination [34]. However as genetic gains attained by conventional breeding decrease, more emphasis should be given to the use of new genetic variability both through pre-breeding or through construction of new parent for crosses, incorporating desired traits from local land races and related wild species, or from other unrelated organism through transgenesis.

The first studies on GE were based on standard variety trials across a series of environments. That allowed identification of the wide or narrow adaptation of the checked cultivars, but little could be said on the genetic basis of adaptation. Extensive field testing of biparental crosses (e.g., [35]), either in the form of doubled haploids, or recombinant inbred lines populations, allows for the assessment of the genetic control of plant adaptation based on standard linkage and QTL analyses, but their use is limited by the level of polymorphisms between parents. In contrast, diverse genotypic panels accumulating multiple recombination events provide ample genetic variation for

association studies. However, their main limitation is the high incidence of false-positive associations due to the difficulty to distinguish between true and pseudo linkage between molecular markers and traits of interest, due to population substructure and correlated selection [36]. More recently, other more complex crossing systems have been proposed to exploit the advantages of both linkage analysis and association mapping. This is the case, for example, of the so-called MAGIC (multiparent advanced generation intercross) [37], the nested association mapping (NAM) design based on a huge set of recombinant inbred lines derived from a large number of founder genotypes [38, 39], and AMPRIL (a multiparent recombinant inbred line population) [40].

The use of physiological criteria in analytical breeding is critical for success [41–44]. Breeders develop a deep knowledge of their target environments and of the agroecological adaptation of their genetic materials. However, whereas intensive work is continuously being carried out by crop physiologists in the area of yield potential and adaptation, not many breeders regularly incorporate new physiological criteria in their mainstream-breeding program. In any case, physiological assessment of adaptation is needed to complement breeders' impressions particularly in the first and last stages of a breeding program: selection of parents and assessment of adaptation of new advanced lines. Similarly, despite exciting progress in molecular marker-assisted selection, applied breeding still depends heavily on direct phenotypic selection of advanced genotypes.

In the rest of this entry, two different aspects will be presented: First, an example of the physiological implications of GE through the study of a trait, time to flowering, that has a clear effect on adaptation; second, a series of increasingly complex statistical models to characterize genotypic adaptation, to identify genotypes showing wide or specific adaptation and to dissect the genetic complexity behind this integrative trait. Although these sections may look quite disconnected, a thorough knowledge of crop physiology and/or their genetic control could allow construction of more powerful integrated statistical models incorporating as genetic covariables this information in order to improve the understanding of the nature of GE. Conversely, the statistical models

can identify certain genotypes which, if well characterized, could allow for empirical identification of key adaptive traits.

Traits Determining Adaptation

The number of physiological traits with a potential effect in determining yield and adaptation is extraordinarily large. In an excellent *Crop Physiology* manual recently edited by Sadras and Calderini [45], many traits are reviewed and organized according to different criteria from capture and efficiency in the use of resources to crop development and plant architecture. Many trade-offs exist between traits that, if ignored, will slow down genetic progress for both potential and actual farmer yields. Araus et al. [43] have also reviewed a number of potentially useful physiological criteria for breeding, particularly, in the framework of C3 cereals under Mediterranean conditions. Crop physiology as a whole is beyond the objectives of this entry. Therefore, the focus is on the single most important crop trait determining plant adaptation, time to flowering, as an example of a key trait to describe the underlying mechanisms and implications for GE.

Time to Flowering

Crop phenology – life cycle as influenced by seasonal variations in climate – has been widely recognized as the most important single factor determining adaptation and thereby crop performance. In determinate species, it allows for matching crop development with availability of resources, avoiding abiotic stresses due to climatic conditions such as late spring frosts and terminal drought. To maximize attainable yield, the most “critical phases” for yield determination have to be matched with the most favorable (or least unfavorable) growing conditions. In some cases (Northern Hemisphere), the obvious way to achieve this is sowing cold-tolerant genotypes early enough to have full growth in early spring, but in the warmer Southern Hemisphere similar maximum yields can be achieved sowing in winter with significantly shorter phases, provided the critical phases are ideally timed [46–49]. Crop phenology is, thus, not only a key adaptive trait, but it may also affect yield potential, since different structures are produced throughout the crop cycle, and some of them may be more important than others in determining

yield potential [50]. If the pattern of water deficit in the target region is relatively predictable, manipulation of genes responsible for crop phenology is the most sustainable approach to increase attainable yield and plant adaptation.

The importance of flowering time has been shown, for example, with the fast and diverse shifts in heading time, or in vernalization and photoperiod responses, due to natural selection: When the same bulk population is grown under contrasting environments [51]; when comparing different sowing dates [52]; when studying the contrasting developmental patterns of genotypes adapted to particular regions [53–55]; or in retrospective studies showing changes in heading date over time due to breeding, particularly in areas where the crop was introduced more recently (e.g., bread wheat in Australia; [48]; durum wheat in certain regions of Spain; [56]). Therefore, crop phenology is an important source of GE for yield when testing genotypes from regions differing in climatic conditions [57, 58].

The three major factors determining flowering time are differential responses to photoperiod and vernalization and intrinsic earliness or earliness per se [50]. Further evidence from recent studies in wheat [59–61] support the idea that earliness per se genes represent basically genotypic differences in the response to non-vernalizing temperatures [62, 63]. The wide genotypic differences for these factors are considered as responsible for the spread of winter cereals, worldwide to a wide range of latitudes and altitudes [49, 64].

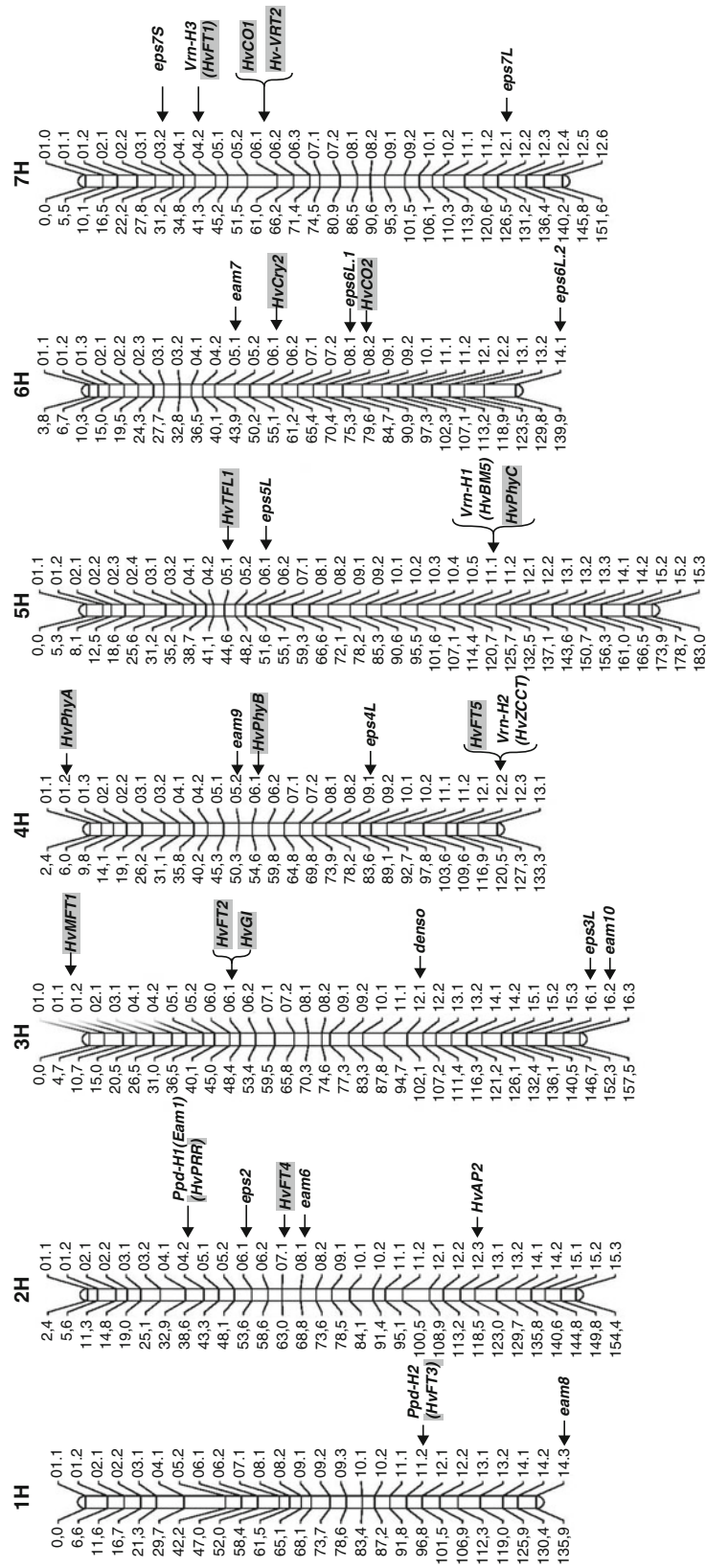
Genetic Factors Controlling Time to Flowering

At the gene or marker level, the importance of flowering time in crop performance is shown, for example, through the geographical distribution of alleles of major genes such as photoperiod (*Ppd*) and vernalization (*VRN*) responsive genes [49, 64, 65]. The co-location of QTLs for heading with QTLs for yield (e.g., [35, 66–69]), which may help to define an optimal window for heading or combination of alleles in the tested environments [70]. Moreover, in some of these studies, QTLs with strong effects on heading collocated with some of the QTLs for yield that exhibited strongest QTL by environment interactions [35, 69–71]. Recent studies have shown, through factorial regression,

that a great part of the effect of these QTLs for heading (underlying QTLxE for yield) can be explained by the different sensitivity of the alleles to environmental conditions such as temperature during different parts of the crop cycle [11, 72].

In the last decade, candidate genes have been identified for major loci controlling flowering time in barley and wheat: The photoperiod responsive gene *Ppd*-H1 in barley and its wheat homologues *Ppd*-D1, *Ppd*-B1, and *Ppd*-A1 are *PRR*-like genes [73, 74]. In both species, the photoperiod-responsive allele accelerates flowering under long-day conditions, but in barley, the greatest differences between sensitive and insensitive alleles are found under long-day conditions or high latitudes, while in wheat, under short day conditions or low latitudes [49, 64, 75, 76]. *HvFT3* is the candidate gene for another gene related to photoperiod in barley, *Ppd*-H2, whose active allele is expressed and accelerates flowering only under short photoperiod or low latitudes [75, 77]. The vernalization genes *VRN*-H1 and its homologues *VRN*-A1, *VRN*-B1, and *VRN*-D1 in wheat are MADS-box transcription factors similar to *APETALA1* in *Arabidopsis* [78–80]. *HvZCCT* and *TaZCCT* are the candidate genes for *VRN*-H2 and its wheat homologue *VRN*-Am2, respectively [81, 82]. The alleles at these loci and their interactions determine the sensitivity to vernalization (e.g., [82, 83]). Finally *VRN*-H3 and its homologues *VRN*-A3, *VRN*-B3, and *VRN*-D3 are *FT*-like genes, which also interact with *PPD* and *VRN* genes [77, 84, 85]. Other reported genes that determine differences in heading time are the “earliness per se” loci (*eps*) identified in barley by Laurie et al. [75], the series of “early maturing” (*Eam*) loci [86–89], and the gene *HvAP2* [90]. However, except for the latter, no candidate genes have been found yet for them and their role is much less clear.

Figure 2 shows the location of the mentioned loci for barley, as well as for some other genes which are homologues to flowering genes in rice and *Arabidopsis* but whose effect on heading is unknown in barley. In wheat, other less characterized loci have also been identified, as the gene *Eps*-2B on 2BS [91, 92]; *Eps*-Am on 1AL sensitive to temperature [59, 60]; *VRN*-D4 close to the centromere in 5D [93], and other earliness per se genes on 5AL [94]. Additionally other loci have been found to have an effect on heading time in different regions than the loci mentioned above, although



Genotype by Environment Interaction and Adaptation. Figure 2

Barley consensus function map showing the location of the vernalization, photoperiod, and earliness per se loci described in the text, as well as some other genes which are homologues to flowering genes in rice and *Arabidopsis*, whose direct effects on heading are still unknown in barley. Distances are given in Kosambi cM and linkage groups are oriented with short arms at the top

most of them with smaller effects: by the use of aneuploids in wheat [49, 95] or through QTL mapping both in barley (e.g., [35, 66–69]) and wheat (e.g., [92, 96–98]). These studies would confirm that heading time is under a strong but complex genetic control [49, 95]. Although particular VRN and PPD alleles may be more frequent in some geographical areas, variation has been found between genotypes within regions, so it is possible finding different combinations of VRN and PPD alleles in successful genotypes well adapted to particular regions, which would reinforce the idea that several other genes may be important in the control of flowering time [64]. As sensitivity to vernalization expresses at earlier stages of development than that to photoperiod, the fact that different combinations of VRN and PPD alleles may confer a similar time to heading or anthesis may also open room for fine-tuning developmental partitioning of a certain time to flowering into different lengths of vegetative and reproductive phases, which might be relevant in improving adaptation (see below).

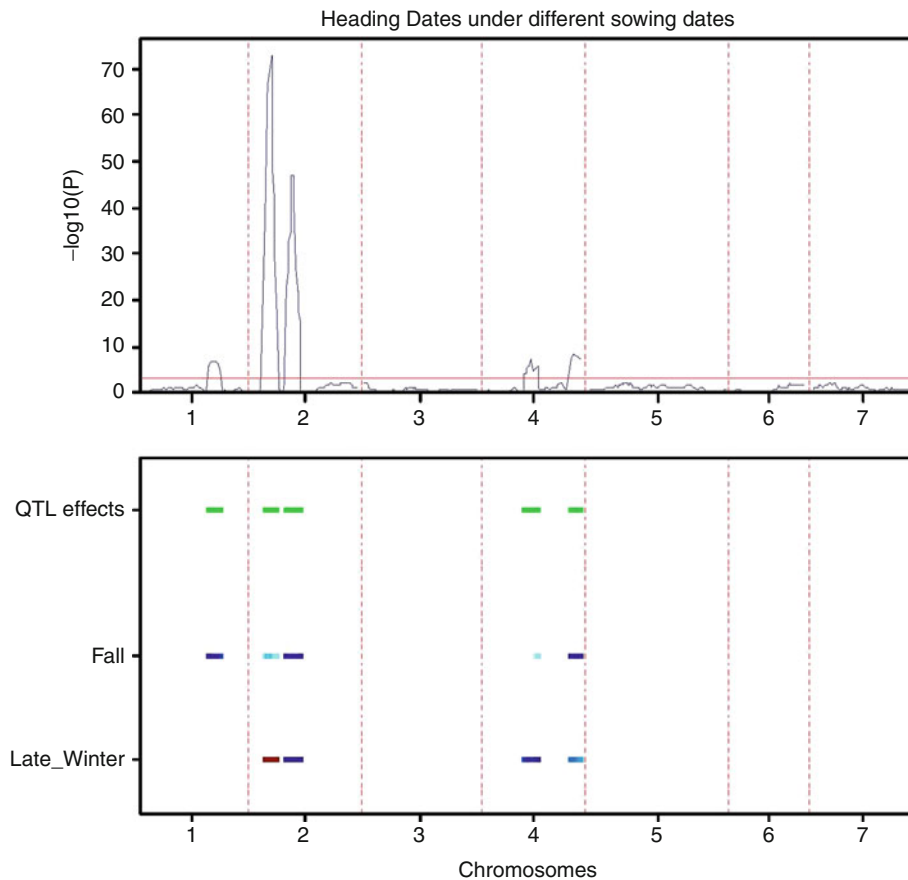
A very simple quantitative genetic analysis of heading date (HD) for the Steptoe \times Morex doubled haploid barley population [35] sown in fall and late winter in 2009 in Spain can be deduced from Fig. 3 which also illustrates alternative types of QTL \times E interactions. In the top part of the figure there is, for a MET situation, a whole genome scan according to a composite interval mapping strategy [99] as implemented by Biometris, Wageningen University and Research Center, in GenStat (version 13th, [100]). All markers in the seven barley chromosomes are represented in sequential order on the X-axis. On the Y-axis is the p value, expressed on a minus logarithmic scale, for the successive regression models, including not just the marker or position of interest, but additional markers that act as cofactors. With $-\log_{10}(p \text{ value})$ increasing, the evidence for a QTL at that position becomes larger. The bottom part of the figure shows firstly, in green, a one-dimensional summary of the profile in the upper panel, that is, all positions for which the joint null hypothesis of no QTL main effect and QTL \times E interaction was rejected. Below the overall test for QTL effects across environments, for each individual environment, in this case defined by fall and late winter planting, an approximate test for environment-specific QTL effects is given in yellow-brown-red (QTL allele second parent

increases trait) or light blue-dark blue (QTL allele first parent increases trait). Two major QTLs seem to determine heading date for the genotypes in these two trials, both on the short arm of Chromosome 2H, corresponding to two known genes, Ppd-H1 and Eam6, on Fig. 2. A very strong qualitative or crossover interaction QTL \times E interaction is shown for Ppd-H1; the Morex allele (yellow-red) in the late winter sowing (under long-day photoperiod) delays heading, whereas the Steptoe allele at this locus (blue) delays heading under short days on the fall sowing. Non-crossover interaction is shown for Eam6. The presence of the Steptoe allele always delays heading, but more under fall sowing (darker blue effect) than under late winter sowing. Other minor QTLs are shown in chromosomes 1H and 4H.

Genetic Factors Controlling Duration of Subphases of Time to Flowering

The effect of these genes or QTLs may vary not only due to different conditions in temperature and photoperiod, or to epistatic interactions with other genes or QTLs, but also they may have different effects on the different phases of the crop cycle. This may be interesting for improving both adaptability and yield potential. Studying the genetic control of different pre-heading phases could bring about a better understanding of crop development patterns and more tools to fine-tuning it. For example, some adaptive characters, such as the avoidance of late frosts in spring, could be better assessed by knowing the duration of the phase from sowing to terminal spikelet rather than total time to anthesis (e.g., [101]). Moreover extending the duration of stem elongation, without modifying total time to anthesis, which is a key trait for adaptability as shown above, has been proposed as a trait to further increase yield potential [102, 103]. This has been proposed because the stem elongation phase is critical for yield determination, as the number of fertile florets at anthesis, which determines the final number of grains, is set during this phase [104, 105].

Several authors have shown that there is partially independent variability between different pre-heading phases (variability in pre-heading phases between genotypes with similar time to heading), both in wheat [106–108] and barley [109–114]. Other authors have



Genotype by Environment Interaction and Adaptation. Figure 3

Genome scan for heading date for the Steptoe \times Morex doubled haploid population grown in fall and late winter sowing in Spain in 2009. *Top:* $-\log_{10}(p)$ values for the test on QTL+QTL.E effects are shown. The red horizontal line indicates the 5% genome-wide significance threshold. *Bottom:* Upper most line in green gives all genomic positions for which null hypothesis of no QTL+QTL.E is rejected. For the fall and late winter sowing environment, all positions for which there is environment-specific QTL expression are indicated with colors: blue showing that the allele from Steptoe delays heading, while red/brown shows that the Morex allele delays heading

shown that responses to vernalization, photoperiod, and temperature can each differ greatly among genotypes and between phases [50, 62, 115, 116]. In some studies using chromosome substitution lines, near isogenic lines and/or single chromosome recombinant lines, hexaploid wheat Ppd-D1 and Ppd-B1 alleles had different effects on the duration of pre-heading phases and on their response to photoperiod, although results seemed to depend on the genetic background and the environmental conditions of each experiment (see results and review by [117]). Recently Lewis et al. [61] found that alleles of a cultivar and a wild line of *Triticum*

monoccocum for Eps-Am had different effects on the leaf initiation and the spikelet initiation phases (due to different sensitivity to temperature), but not on stem elongation, while they had little effect on total time to heading. On the other hand, many of the QTLs responsible for a different genetic control between pre-heading phases had little or no effect on total time to heading, so they may be more difficult to detect when assessing only heading time [111, 118]. Some of these differences in the length of pre- and post-heading phases were maintained under different conditions of photoperiod and temperature [119].

Statistical Approaches for GE Characterization

Means across environments are adequate indicators of genotypic performance only in the absence of crossover GE. When present, the use of means across environments ignores the differential reaction of genotypes to environmental changes. In an analysis of variance, introduction of the GE interaction term, $(GE)_{ij}$ for $i = 1$ to g genotypes and $j = 1$ to e environments, creates as many parameters as there are GE combinations, making predictions of phenotypic responses for environments that were not in the set of trial environments impossible. Most approaches for the study of GE interaction and adaptation depart from ANOVA models with GE interaction terms and are therefore purely empirical descriptions of phenotypic performances of a set of genotypes across a fixed sample of environments. However, if the physiological or environmental underlying causes determining GE interaction can be determined, identification of genotypes better adapted to certain specific environmental conditions would be possible and, thus, larger genetic gains would be achievable. Furthermore, if the traits conferring adaptation and their genetic control are revealed, direct implementation in breeding may be feasible.

This entry reviews three types of statistical approaches used in GE interaction for breeding and variety development: (1) regression on the environmental mean, best known as Finlay–Wilkinson regression, or joint regression analysis; (2) linear-bilinear models, like AMMI and GGE; and (3) factorial regression models (see specific references for these methods below). These methods differ not only on the information they provide, but also in their predictive ability for breeding. A discussion of these three types of models from a common statistical perspective can be found in [120, 121]. The approaches aim at substituting the $(GE)_{ij}$ term by a linear or bilinear approximation using fewer parameters (Table 1). The replacement of double-indexed ANOVA GE interaction parameters by single-indexed regression and bilinear parameters introduces predictive properties.

Regression on the Mean

The most widely used and abused statistical method in breeding programs for characterizing GE has been the

regression-on-the-mean analysis first proposed by Yates and Cochran [122] and made popular by Finlay and Wilkinson [123] (FW), and also named joint regression analysis. This method summarizes phenotypic responses to environmental changes as straight lines differing in both intercept (related to genotypic main effect) and slope (which estimates environmental sensitivity); GE interaction is revealed by differences in the slopes of individual genotypes. These straight lines are produced upon regressing individual genotypic means per environment on average site performance across all genotypes in that environment, where the regression is done across the full set of environments.

The rationale behind FW is that in the absence of explicit environmental information, a good estimate of the agronomical value of any environment may be given by the average phenotypic performance of all genotypes in that environment. This method has an important conceptual drawback. Two environments may have a similar low average yield for two completely different agroecological reasons, for example, presence of a disease and an episode of a late spring frost just before flowering. This model assumes the genotypic sensitivity to these two stresses to be approximately the same when the different stresses produce the same environmental means. Therefore, the use of the model is best restricted to those rare cases in which environmental differences are driven by just a single major biotic or abiotic factor; in these cases, the linear regression on the mean model may reflect linear differences in relation to the predominant stress factor. However, if environmental differences are due to a major stress, why not using, rather than the average phenotypic value at every environment, a direct estimate of the genotypic sensitivity to this stress as in the factorial regression method described below?

Regression-on-the-mean models are conceptually simple: The differential genotypic responses are summarized by their slopes, but it is very important to point out that their value and use should depend on the proportion of GE sum of squares that can be described by the differential environmental sensitivities of the genotypes. Figure 4 presents an example for which the Finlay and Wilkinson model should have never been used; however, it has been presented in this entry as similar reports are still too often seen in

Genotype by Environment Interaction and Adaptation. Table 1 Overview of statistical models for GE analyses from two-way genotype by environment table of means derived from MET

General model	Specific model	Model	Data required	Statistical models for $E(Y_{ij}) - \mu$	Key information provided ^a
Reference models	<i>Additive</i>	I	Phenotypic data ^b	$G_i + E_j + e_{ij}$	Average cultivar yields
	<i>Full interaction</i>	II	Phenotypic data	$G_i + E_j + (GE)_{ij}$	Departures from additivity for each environment
Regression on the mean	<i>Finlay and Wilkinson</i>	III	Phenotypic data	$G_i + E_j + \beta_i E_j + e_{ij}$	Cultivar sensitivity (in form of slopes) to changes in environmental productivity
Bilinear models	<i>AMMI</i>	IV	Phenotypic data	$G_i + E_j + \sum_{k=1}^K a_{ki} b_{kj} + e_{ij}$	Joint adaptation patterns of genotypes to environments
	<i>GGE</i>	V	Phenotypic data	$E_j + \sum_{k=1}^K a'_{ki} b'_{kj} + e_{ij}$	Identification of the “winning genotype” for each uniform subset of environments
Factorial regression models	<i>Factorial regression model</i>	VI	Phenotypic and environmental data	$G_i + E_j + \beta_i z_j + e_{ij}$	Cultivar sensitivities (β_i) to changes in any environmental variable z
	<i>Genotypic factorial regression model: QTLE model</i>	VII	Phenotypic and genotypic (marker information) data	$x_i \rho + E_j + x_i \rho_j + e_{ij}$	Marker (x) potentially associated to QTL and to QTLE and the corresponding QTL (ρ) and the QTLE (ρ_j) effects ^c
	<i>Integrated factorial regression model</i>	VIII	Phenotypic, genotypic, and environmental data	$x_i \rho + E_j + x_i (\lambda z_j) + e_{ij}$	QTL sensitivity to changes in environmental variable z ^d

^aSee text for a more detailed discussion of each model

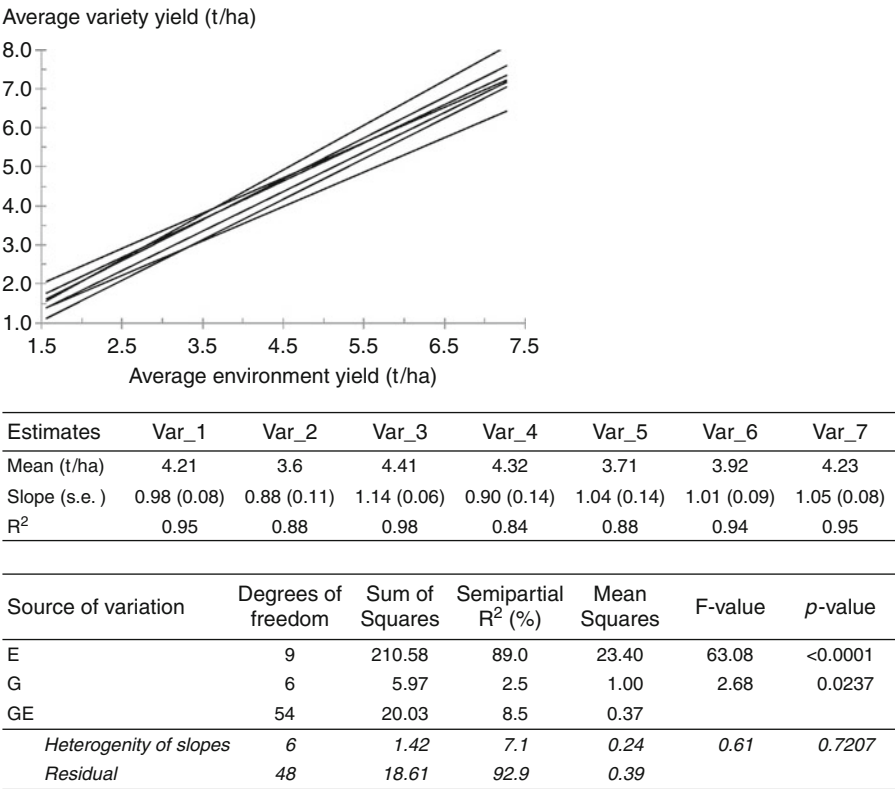
^bPhenotypic response of the $i = 1 \dots g$ genotype at the $j = 1 \dots e$ environment

^cIn the presence of QTLE, ρ_j adjusts the average QTL expression across environments, ρ , to a more appropriate level for the individual environment j . This model can be easily extended to x_s markers throughout all the genome

^d λ is a constant that determines the extent to which a unit change in z , an environmental covariable, influences the effect of a QTL allele substitution. This model can be easily extended to x_s markers and z_t environmental variables

many publications. It summarizes a small MET consisting of seven barley varieties (Var_1 to Var_7) grown at ten Spanish environments according to model III in Table 1. In the part of this figure, there are the simple linear regression models for the seven varieties. If nothing else is shown, it can be wrongly assumed that there are substantial differences among genotypic slopes. This is also shown on the top table that includes regression estimates. When independent simple linear regression analyses are fitted for the seven genotypes, the slopes varied from 0.88 to 1.14 and the individual

straight lines were very significant (R^2 from 84% to 98%; p values from 1.8×10^{-04} to 7.1×10^{-08}). However, these R^2 s do not mean anything in the GE context. They simply confirm that the genotypic yield increases with the mean environmental yield, which is obvious in the way that this model is built. Based on these estimates, it can be wrongly stated, for example, that Var_3 (slope equal to 1.14) apparently benefits more to improvements in the overall productivity of the environment than Var_6 (1.01) and particularly than Var_2 (0.88) which, with the lowest sensibility,



Genotype by Environment Interaction and Adaptation. Figure 4
Inappropriate use of the Finlay and Wilkinson analysis for a MET consisting of seven barley genotypes grown in ten environments in Spain

does worst than expected. However, this model is completely inadequate for this MET and the previous estimates are useless and misleading and should have never been determined. The standard errors of the slopes, which can be used to assess the significance of the differences among slopes, ranges from 0.06 to 0.14, with an average standard error of the difference equal to 0.16. They are too large for detecting significant differences between genotypic slopes. Furthermore, joint regression analysis of variance table (bottom part of Fig. 4) shows that the observed differences among the genotypic slopes (Heterogeneity of slopes) only explains 7.1% of the GE sum of squares, which is not statistically significant (p value = 0.721).

Bilinear Models (AMMI and GGE)

The usefulness of the integration of ecophysiological and statistical tools in the interpretation of GE interaction is

examined based upon the joint application of two multiplicative models for interaction: the additive main effects and multiplicative interaction (AMMI) model [6], and the factorial regression model [120, 124]. Both provide information and insight beyond the classical analysis of variance of two-way genotype by environment tables. AMMI represents an empirical approach (based on yield itself) to analyze GE interaction. Factorial regression attempts to describe interaction by including external genetic, phenotypic, and environmental information (e.g., morphophysiological traits, climatic data, etc.) on the levels of the genotypic and environmental factors. It implies a more analytical approach to the understanding of GE.

The Finlay and Wilkinson model belongs to a wider class of statistical models named linear-bilinear which estimate genotypic sensitivities to one or more environmental characterizations that are just linear functions of the phenotypic data [124–127]. However, the

additive main effects and multiplicative interaction (AMMI) model [128–131] and the GGE models [132, 133] represent more powerful, and thus, useful examples of linear-bilinear models in plant breeding. These two model classes generate for every genotype and for every environment a series of K scores, which summarize the differential sensitivity of the genotypes to the prevalent, and typically unknown, stresses present in the analyzed MET.

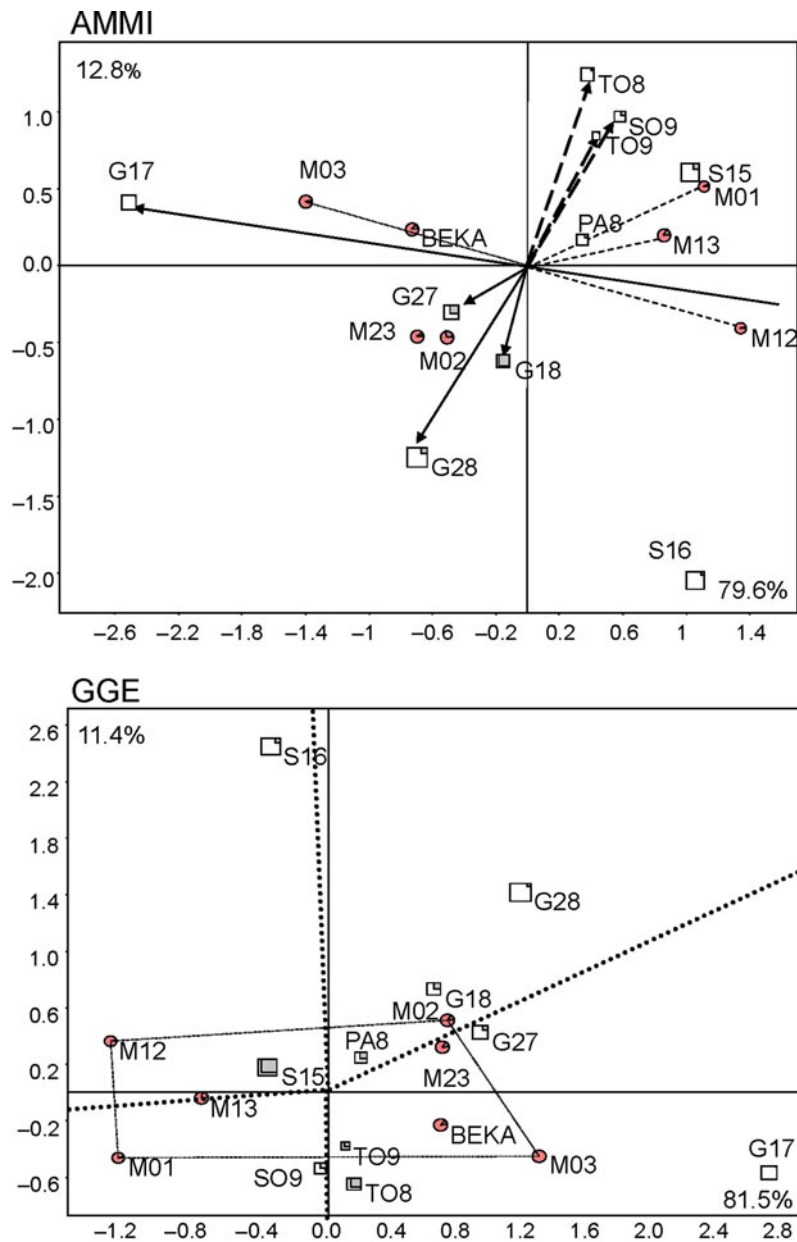
The AMMI model successively partitions the $(GE)_{ij}$ interaction term from the basic ANOVA reference model into a series of K multiplicative terms or products of the form $a_{ki}b_{kj}$ where, for the k th term, a_{ki} refers to the genotypic sensitivity of genotype i to an hypothetical environmental variable b_k , which has value b_{kj} in environment j (Table 1, model IV). Alternatively, b_{kj} refers also to the environmental potentiality of environment j to an hypothetical genotypic variable a_k , which takes value a_{ki} for genotype i . The K hypothetical environmental (genotypic) variables have the property of discriminating maximally between genotypes (environments). The number of multiplicative terms to be retained for an appropriate estimate of the GE interaction, K , can be estimated in various ways, see, for example, Gollob [130], Gauch [6], and Cornelius [134]. From a practical point of view, the AMMI model is fitted in two steps. First, an additive ANOVA model is fitted containing the main effects for G and E and then the residuals from the additive model are used to construct the GE interaction matrix. This interaction matrix is then subjected to a singular value decomposition that generates the above-introduced genotypic and environmental scores [128, 130, 131].

Key outputs of the AMMI analysis are the genotypic and environmental scores for the K retained axes, along with the proportions of the interaction sum of squares explained by the multiplicative terms. The output of the $K = 2$ AMMI model, retaining just the first two interaction axes (IPCA1 and IPCA2), can be directly visualized by means of a biplot [5, 128, 135]. If both axes together explain most of the GE interaction, interpretation of the biplot is very simple and potentially extremely useful for understanding GE interaction. The i th genotype is placed in the biplot according to the (a_{1i}, a_{2i}) genotypic scores; similarly, the j th environment is defined by its two IPCA environmental scores (b_{1j}, b_{2j}) . Distance

of a genotype or environment to the origin is proportional to the GE interaction generated by that genotype or environment, respectively. Genotypes placed close together show similar adaptation patterns. Close environments generate similar GE interactions.

The actual interaction of genotype i in environment j can be estimated by the projection of the genotype position (a_{1i}, a_{2i}) on the j th environmental vector that goes from the origin $(0,0)$ to (b_{1j}, b_{2j}) , that is the line that goes through the origin with slope equal to b_{2j}/b_{1j} . The distance between the genotype projection on the line to the origin also provides information about the absolute magnitude of the interaction of genotype i in environment j . Genotype i will be well adapted to environment j , that is, positive interaction, if the projection is in the direction of the environmental vector and negative otherwise. The sign of the interaction of the genotype i in environment j can be estimated by the cosine between the i th genotypic and the j th environmental vector. It will be positive if both vectors form acute (close to 0°) angles, negative if the angle is obtuse (close to 180°), and nonexistent (no interaction) if they form a right angle (close to 90°). In a similar way, two environments whose vectors form an acute angle generate a similar type of GE interaction across genotypes, the environments have positive genetic correlation. If the two environmental vectors form an angle close to 180° , whichever genotype is well adapted in one environment will be poorly adapted to the other, the environments have a negative environmental interaction. Finally if both environmental vectors form a right angle, the genotypic behavior at one environment will be independent of the behavior at the other site, the genetic correlation is zero.

The upper part of Fig. 5 shows an AMMI biplot generated by a set of seven genotypes grown at ten environments. The genotypes are shown by circles and they represent a barley variety Beka, three derived single nonallelic mutants, M01, M02, M03, and the three binary mutant combinations, M12, M13, M23. The environments are shown in the biplot by squares which represent location by year combinations across Spain. Production of these mutants and analysis of these data was presented elsewhere [136, 137]. In this MET, the GE interaction is well described by the AMMI $K = 2$ model, as both axes explain together more than



Genotype by Environment Interaction and Adaptation. Figure 5
AMMI and GGE biplots for a MET consisting of seven barley genotypes grown at ten environments in Spain (Data taken from [105]). See text for a detailed description of genotypes and environments

90% of the GE sum of squares. The average yield of each environment and genotype is shown proportional to the area of its corresponding symbol. Within each symbol there is a, generally small, darker sector that represents the proportion of its sum of squares not explained by this model. In this case all environments are well represented except for G27 y G18, which generate GE interactions not correctly described by the

AMMI $K = 2$ model. Beka is placed close to the origin and, thus, it is the genotype that interacts least with the ten environments; on the contrary, M12 and M03 are the two genotypes that interact most with the environments. G17 and S16 are the two environments which showed the largest GE interaction, that is, whose genotypic yields depart most from their averages. PA8, near the origin, produced yields close to the average across all environments.

The relative position of both genotypes and environments can provide some clarification on the nature of the GE interactions in this MET. The first IPCA seems to be associated with differential behavior of genotypes carrying the first mutation, M01, M12, and M13, with positive scores in comparison to the other genotypes. These mutants are particularly poorly adapted to Granada (G in the biplot, especially G17). The second axis, which is quantitatively less important, seems associated with mutant 2 (M02, M12 y M23), which shows negative scores on this axis, whereas the other genotypes have positive scores; the specific adaptation of this mutant to the environments is not as clear.

The angle formed by any two environmental vectors is related to the relative similarity among environments, say, the genetic correlation, as determined by the genotypic yields. In this case, the relative yields of the genotypes in Toledo (TO8 y TO9) seem very similar to Soria (SO8). They all form acute angles with cosine and correlation close to 1. T09, with a smaller size square, had lower yields than the others. By comparing the angle of these three environmental vectors with the vector determined by G28 (very obtuse angle closed to 180° and cosine and correlation close to -1), it can be deduced that those genotypes that behave relatively well in G28 perform poorly in the other three sites and vice versa. The analysis of the genotypic projection on environmental vectors gives clues about specific adaptation patterns. For example for G17, M03 showed a good adaptation to this environment, whereas M12 was particularly poorly adapted there. This AMMI analysis was done on the MET data used for the analysis in Fig. 4. Whereas the Finlay and Wilkinson method was able to explain only 7% of the GE sum of squares, the AMMI model for $K = 2$ retained 90% of the GE sum of squares. Furthermore, as described in the previous paragraphs, the known structure of the seven genotypes

developed through artificial mutagenesis, suggested a model with a plausible genetic meaning.

The environmental and genetic scores are simple statistical estimates derived from MET phenotypic data, without any direct physiological meaning. However, these empirical estimates can be associated to physiological processes by correlating the environmental scores to explicit environmental measurements, such as soil or meteorological variables; these correlations can often provide meaningful agroecological information about the nature of GE interactions [11, 14, 138–140].

Another member of the linear-bilinear model class is the GGE model [132, 133], in which single value decomposition is done on the sum of the G and GE components by just subtracting the environmental means (environmental centered) on the two-way table of means (Table 1, model V) rather than on GE interactions alone, as done in AMMI. A GGE biplot for $K = 2$ provides additional information of potential interest to breeders, as it allows for the direct identification of the “winning” genotype in any potentially uniform subset of environments. To do so, the most extreme genotypic scores are connected delimiting an irregular polygon enclosing all other genotypes, that is, a convex hull is constructed. In the previous example (Fig. 5, bottom) this is an irregular quadrilateral defined by M12, M02, M03 y M01. Next, lines perpendicular to each side of the polygon/convex hull are drawn (thicker lines in Fig. 5, bottom) up to the boundaries of the biplot. In this way sectors are created, called mega environments, which contain environments that behave relatively uniform with respect to the genotypes. The “winning genotype” in a mega environment is the genotype that is placed at the vertex of the polygon inside that mega environment. For example, M12 is the best-adapted genotype in the mega environment defined by S15 and, particularly, S16. Mutant M03 is the most productive genotype in G18 and G28. Of course, this interpretation is subjected to the condition that most of G+GE variability is retained in the first two GGE axes.

Factorial Regression Models

Factorial regression models were developed to incorporate additional explicit environmental information

(variable z in Table 1 model VI) into a model [120, 121] for GE interaction and estimate the genotypic sensitivity of each of g genotypes (β_i in Table 1 model VI) to these independent variables (regressors, covariables). The regression on the mean or FW analyses reported before may be seen as a specific case of factorial regression, in which the average yield in each environment is used as an explicit environmental characterization. In the general form, any explicit agroecological variable individually recorded for each environment could be used as independent explanatory variable. Average yield can be a reflection of a certain meteorological variable, such as available soil water. In this situation, this variable recorded for each environment could be used as explanatory independent variable to describe GE interaction (variable z in model VI Table 1). The genotypic slopes will have a more direct physiological meaning when they estimate, for example, sensitivity to changes in available soil water, which is an approximation to water use efficiency. In a triticale MET, GE interaction for grain yield was regressed on soil pH and the genotypic slopes directly assessed the sensitivities to changes in soil pH [141]. Extension to multiple environmental variables and complex response curves is conceptually simple and easily computable using standard statistical packages. As for any multiple regression models, a central question is the choice of variables for description of GE interaction. Continuous monitoring of the environment generates huge numbers of environmental covariables, which will complicate identification of the most relevant ones. Purely statistical selection procedures often lead to physiologically incomprehensible models. Therefore, agroecological insights of genotypes and environments should augment and prevail over purely statistical considerations. A helpful prescreening of environmental covariables can sometimes be done by correlating covariables to scores derived from AMMI or GGE analyses [11].

Factorial Regression Models Incorporating Explicit Genotypic Information Genotypic covariables can also be used to partition the G and GE terms. Molecular markers such as DNA polymorphisms for anonymous sequences or for functional genes are the most useful and readily available genetic covariables. For a codominant marker in a diploid species with

potential genotypes AA , Aa , and aa , the number of A alleles (2, 1, and 0 to represent genotypes AA , Aa , and aa , respectively) could be used as a genetic covariable, x , in a factorial regression model (Table 1, model VII). If multiple markers across the whole genome are sequentially used, factorial regression has the ability to detect, locate, and estimate QTL main effects and QTL by environment interactions. For marker positions adjacent to a QTL, the ρ slope in model VII (Table 1) estimates directly the effect of a QTL allele substitution. Similarly, the $(GE)_{ij}$ interaction can be further partitioned into a term for differential QTL expression across environments, ρ_p , and a residual GE interaction. For a full genome scan, factorial regression models can be fitted on grid of genomic positions, on markers and in between markers, when necessary. Virtual markers, in between observed markers, can be easily generated from flanking marker information (see [142]). Factorial regression models which include genetic covariables can be potentially used for any set of genotypes for which genetic predictors can be constructed, from standard biparental offspring populations and unrelated diverse association panels, to more complicated intercross systems, such as MAGIC [37], NAM [38, 39], and AMPRIL [40] described before. The QTL.E interaction model shown in model VII (Table 1) is based on application of a simple marker regression to our data. To construct multiple QTL models, a composite interval mapping approach can be followed by incorporating cofactors, or markers that correct for QTL elsewhere, on the genome.

Factorial Regression Models Incorporating Explicit Environmental and Genotypic Information The final goal of any MET is to understand the nature of GE interaction in terms of differential sensitivity of the different QTLs or genes to external environmental variables. This is also possible by means of factorial regression models [11, 13, 72, 92]. Differential QTL expression for environments, ρ_p , can be regressed on any environmental covariable, z , to relate the differential QTL expression directly to key environmental variables responsible for GE. This is done by substituting the QTL.E term, $x_i\rho_p$, with a linear regression $x_i(\lambda z_j)$ and a residual term. λ is a constant that determines the extent to which a unit change in z , the environmental

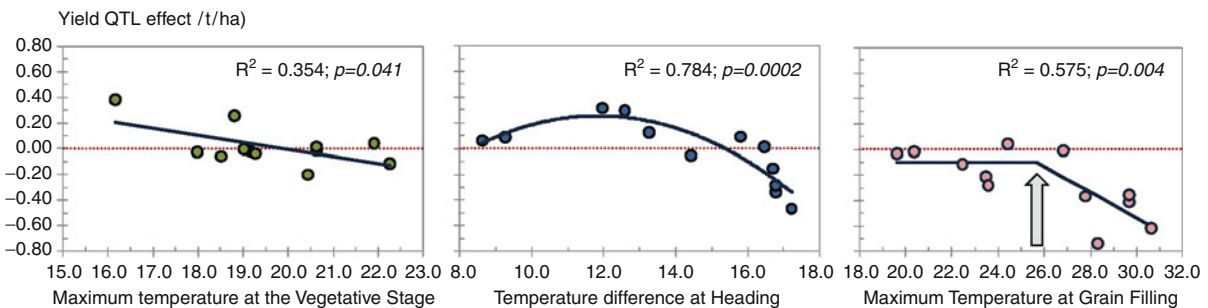
covariable, influences the effect of a QTL allele substitution. The statistical model used is listed as model VIII in Table 1 which can be easily extended to multiple markers (x_{si}) and various environmental variables (z_{tj}).

Van Eeuwijk et al. [143, 144] and Boer et al. [99] provide examples of differential QTL expression in maize data to environmental variables; by incorporating marker information and environmental covariables describing the environment, these models allow for prediction of differential genotypic sensitivities to environmental changes. An example of the output of these fully integrated genotypic and environmental models is shown in Fig. 6, which shows an analysis for the “Steptoe \times Morex” double haploid population data from the North American Barley Genome Project, grown at 12 sites and with environmental covariables at hand. Three main QTLs were responsible of GE interaction [71]. Differential QTL effects across environments could be associated to three different environmental variables related to temperature taken at three different growth periods and according to three alternative models: a simple linear regression model, a second degree response, and a “broken-stick” model (Fig. 6). Furthermore, two out of three QTL.E interactions showed a “crossover” type interaction: The sign of the QTL effect changed according to the value of the environmental external variable. This

figure clearly illustrates the importance of QTL.E interaction for complex traits such as grain yield in barley.

The Mixed Model Framework: Modeling Variance-Covariance Structures

Table 1 shows different alternatives for modeling the expected responses of a genotype to environmental changes, without any specific concern about the implicit assumptions of the analyses of variance. Standard linear models take for granted that error terms are independent and have constant variance. However for MET, these assumptions are overly simplistic as variances within environments and correlations between environments tend to be heterogeneous. For the sake of brevity and simplicity, how the mixed model framework also allows for modeling of the variance-covariance component of the data has not been described. However, the optimal statistical modeling for MET data should focus first in finding an adequate variance-covariance model for the random terms and then, as discussed above, search for a parsimonious model for the expected responses. Choice of variance-covariance model can have strong implications. In the case of QTL modeling, QTL may erroneously be declared significant or nonsignificant because of over or under estimation of effect sizes and standard errors [72, 145]. The mixed model framework, which combines modeling of means and variances, provides



Genotype by Environment Interaction and Adaptation. Figure 6

Differential sensitivities of three major QTLs to temperature, recorded at three different growing periods for the Steptoe \times Morex doubled haploid population (Data taken from the North American Barley Genome Project). Twelve sites with environmental characterizations were available. Three different models were used: a straight-line regression model, a second-degree polynomial, and a “broken-stick” factorial regression model

a more appropriate modeling environment for GE and QTL.E interactions offering flexibility with regard to assumptions on heterogeneity of variances and on correlations across environments [17].

Computer Software for GE Analyses

Annicchiarico [3] lists a series of user-friendly computer software available for many GE analyses. CROPSTAT is a freely available package developed by the International Rice Research Institute [146] that has specific modules for FW and AMMI analysis. MATMODEL available in a free version [147] also provides AMMI and joint regression modeling and it is particularly useful for handling missing data. INFOGEN [148] within the INFOSTAT system [149] also includes most described tools for the analysis of MET. At the same time, there are also dedicated commercial softwares, such as GGE BIPLLOT [132], useful for joint regression, AMMI, and GGE. Obviously, all general statistical packages can easily be programmed to fit all linear-bilinear models described in this entry in a fixed model context, whereas some like GenStat, ASREML, and SAS also allow fitting mixed bilinear models. SAS instructions for many GE analyses are presented in Kang [100]. GenStat [150] includes specific procedures for FW, AMMI, and GGE analyses. Version 13 of GenStat (2010) also includes dedicated menus for QTL and QTL.E analyses for segregating crosses and for association analyses. GenStat has a policy of free licensing of older versions to institutions in developing countries and for educational purposes in the form of the GenStat Discovery version.

Future Directions

Plant breeding research experiences fast changes. Nowadays, at the genomic side, sequencing and single-nucleotide polymorphism (SNP) technology is becoming increasingly cheap for not only model species, but also for crop species. Besides information at the DNA level, genomic information at RNA, protein, and metabolite level starts to become common. As a consequence, huge amounts of data start to become available for characterizing genotypes at various genomic levels. Similar developments can be observed for monitoring the environment. Environmental

characterizations can be stored over the growing season for all environmental factors that are believed to be relevant.

In the past, genotypic and environmental information was the bottleneck; however, the current focus has shifted to access to the right plant material and their correct phenotyping. High-throughput phenotyping techniques are being developed that facilitate monitoring of individual plants at arbitrary small intervals over the growing season. However, high-throughput phenotyping schemes taken in individual cell/tissue/organ/single plant level may not mean anything at the crop level. Up-scaling from processes taking place in a fraction of a second and in a fraction of space to relevant crop traits (produced in a hectare through several months) has consistently failed in the past and remains a challenge. Crop physiology can play a key role in understanding multi-trait interactions for up-scaling from gene to crop.

The strongly increased availability of phenotypic, genomic, and environmental information begs for new statistical techniques that allow the increased information to be used in an effective way. Various requirements can be defined. First, phenotypic information will increasingly concern a wide array of traits that are repeatedly measured over time. Correlations between these traits will need to be explicitly modeled, as will be the correlations between the repeated measurements for the same trait. Information from multiple environments can be treated in the same way as information from multiple traits, although correlations between the same trait in different environments may ask for other models than the correlations between different traits in the same environment and different traits in different environments. Standard mixed model procedures will fail, as too many variances and covariances/correlations will require estimation. A way out may be to regularize the pattern of variances and covariances by inserting biological information in the estimation in the form of alternative statistical tools, such as priors (Bayesian methods) or penalties (penalized multivariate regressions). One popular way of reducing the number of correlation parameters is by imposing network structures on sets of trait by environment combinations, thereby effectively fitting sparse matrices to the inverses of the correlation matrices. The graphical lasso is an example of such an approach [151].

Turning to increased marker numbers and selecting meaningful genotype to phenotype models in the face of 100,000s of SNP markers demand new statistical approaches. As identification of individually contributing SNPs in such conditions is very difficult, an alternative strategy emphasizing prediction from markers above identification of markers is rapidly gaining popularity. In genomic selection, the idea is to use all markers simultaneously for predicting marker-based breeding values that help in ranking individuals on genetic merit [152–154]. Bayesian and penalized regression techniques help to regularize the estimates for individual marker contributions, as it will be evident that with standard regression techniques it is impossible to estimate hundreds of thousands of marker effects. Mixed models can in this context be interpreted as an example of a Bayesian technique in which the prior for the marker effects is a normal distribution. Equivalently, mixed models can be seen as penalized regressions in which the ratios of variance components determine the penalties (shrinkage factors). As an illustration, one may regress a phenotypic trait on a large set of markers, assuming the effects of the markers of individual chromosomes to follow normal distributions with chromosome-specific variances. The predicted values for the genotypes from such a mixed model represent the genomic breeding value. This breeding value can be used for selection purposes. Examples of genomic selection for multiple environments are still hard to find.

The increased information from intensive environmental monitoring can be used to improve prediction of genotypic performance by integrating it with other types of genotype-specific information in crop growth models [16, 17, 155–157]. The environmental information is fed into a suitable crop growth model and when physiological parameters of the crop growth model can be specified at genotype-specific level, the crop growth model can produce predictions for individual genotypes in any environment for which a full environmental characterization is given. An integration of crop growth modeling with genomic selection is possible when the values for the genotype-specific physiological parameters in the crop growth model are inserted from Bayesian or mixed genomic selection models.

The increased amounts of phenotypic, genomic, and environmental data pose strong demands on our

statistical ingenuity, but interesting solutions start to appear on the horizon. In this forthcoming scenario, elaborations of mixed models, Bayesian techniques and penalized methods will play a major role in the analysis of GE interactions.

Bibliography

Primary Literature

1. Student (1908) The probable error of a mean. *Biometrika* 6:1–25
2. Annicchiarico P (2002) Genotype \times environment interactions: challenges and opportunities for plant breeding and cultivar recommendations. FAO plant production and protection paper no. 174. FAO, Rome
3. Annicchiarico P (2009) Coping with and exploiting genotype-by-environment interactions. In: Ceccarelli S, Guimaraes EP, Welzien E (eds) Participatory plant breeding. FAO, Rome, pp 519–564
4. Cooper M, Hammer GL (eds) (1996) Plant adaptation and crop improvement. CAB International, Wallingford
5. Fox PN, Crossa J, Romagosa I (1997) Multi-environment testing and genotype by environment interaction. In: Kempton RA, Fox PN (eds) Statistical methods for plant variety evaluation. Chapman and Hall, London, pp 117–137
6. Gauch HG (1992) Statistical analysis of regional yield trials. Elsevier, Amsterdam
7. Kang MS (1998) Using genotype-by-environment interaction for crop cultivar development. *Adv Agron* 62:199–252
8. Kang MS, Gauch HG (1996) Genotype by environment interaction: new perspectives. CRC Press, Boca Raton
9. Kempton RA, Fox PN (1997) Statistical methods for plant variety evaluation. Chapman and Hall, London
10. Romagosa I, Fox PN (1993) Genotype-environment interaction and adaptation. In: Hayward MD, Bosemark NO, Romagosa I (eds) Plant breeding, principles and prospects. Chapman and Hall, London, pp 373–390
11. Romagosa I, van Eeuwijk FA, Thomas WTB (2009) Statistical analyses of genotype by environment data. In: Carena M (ed) Handbook of plant breeding, vol 3, Cereals. Springer, New York, pp 291–331
12. van Eeuwijk FA (2006) Genotype by environment interaction: basics and beyond. In: Lamkey K, Lee M (eds) Plant breeding: the Arnell Hallauer international symposium. Blackwell, Oxford, pp 155–170
13. van Eeuwijk FA, Malosetti M, Yin X, Struik PC, Stam P (2005) Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Aust J Agr Res* 56:883–894
14. Voltas J, van Eeuwijk FA, Igartua E, Garcia del Moral LF, Molina-Cano JL, Romagosa I (2002) Genotype by environment interaction and adaptation in barley breeding: basic concepts and methods of analysis. In: Slafer GA, Molina-Cano JL, Savin R,

- Araus JL, Romagosa I (eds) Barley science: recent advances from molecular biology to agronomy of yield and quality. Haworth Pres, Binghamton, pp 205–241
15. Paterson AH (1998) Molecular dissection of complex traits. CRC Press, Boca Raton
 16. Cooper M, van Eeuwijk FA, Hammer GL, Podlich DW, Messina C (2009) Modeling QTL for complex traits: detection and context for plant breeding. *Curr Opin Plant Biol* 12:231–240
 17. van Eeuwijk FA, Bink MCAM, Chenu K, Chapman SC (2010) Detection and use of QTL for complex traits in multiple environments. *Curr Opin Plant Biol* 13:193–205
 18. Calderini DF, Slafer GA (1999) Has yield stability changed with genetic improvement of wheat yield? *Euphytica* 107:51–59
 19. Foulkes MJ, Sylvester-Bradley R, Weightman R, Snape J (2007) Identifying physiological traits associated with improved drought resistance in winter wheat. *Field Crop Res* 103:11–24
 20. Reynolds MP, Borlaug NE (2006) Impacts of breeding on international collaborative wheat improvement. *J Agric Sci* 144:3–17, Cambridge
 21. Slafer GA, Araus JL (2007) Physiological traits for improving wheat yield under a wide range of conditions. In: Spiertz JHJ, Struik PC, van Laar HH (eds) Scale and complexity in plant systems research: gene-plant-crop relations. Springer, Dordrecht, pp 147–156
 22. Fischer RA, Edmeades GO (2010) Breeding and cereal yield progress. *Crop Sci* 50:S85–S98
 23. Reynolds MP, Foulkes J, Slafer GA, Berry P, Parry MJ, Snape JW, Angus WJ (2009) Raising yield potential in wheat. *J Exp Bot* 60:1899–1918
 24. Slafer GA (2003) Genetic basis of yield as viewed from a crop physiologist's perspective. *Ann Appl Biol* 142:117–128
 25. Siddique KHM, Belford RK, Perry MW, Tennant D (1989) Growth, development and light interception of old and modern wheat cultivars in a Mediterranean environment. *Aust J Agr Res* 40:473–487
 26. Slafer GA, Andrade FH (1993) Physiological attributes related to the generation of grain yield in bread wheat cultivars released at different eras. *Field Crop Res* 31:351–367
 27. Miralles DF, Katz SD, Colloca A, Slafer GA (1998) Floret development in near isogenic wheat lines differing in plant height. *Field Crop Res* 59:21–30
 28. Miralles DJ, Slafer GA (1995) Yield, biomass and yield components in dwarf, semidwarf and tall isogenic lines of spring wheat under recommended and late sowing dates. *Plant Breed* 114:392–396
 29. Richards RA (1996) Increasing yield potential in wheat – source and sink limitations. In: Reynolds MP, Rajaram S, McNab A (eds) Increasing yield potential in wheat: breaking the barriers. CIMMYT, Mexico, pp 134–149
 30. Foulkes J, Slafer GA, Davies WJ, Berry P, Sylvester-Bradley R, Martre P, Calderini DF, Griffiths S, Reynolds M (2011) Raising yield potential of wheat. III. Optimizing partitioning to grain while maintaining lodging resistance. *J Exp Bot* 62:469–486
 31. Denison RF (2009) Darwinian agriculture: real, imaginary and complex trade-offs as constraints and opportunities. In: Sadras VO, Calderini D (eds) Crop physiology. Applications for genetic improvement and agronomy. Academic, Burlington, pp 215–234
 32. Grando S, Ceccarelli S (2009) Breeding for quantitative variables. Part 3: breeding for resistance to abiotic stress. In: Ceccarelli S, Guimaraes EP, Weltzien E (eds) Participatory plant breeding. FAO, Rome, pp 391–417
 33. Cooper M, Fox PN (1996) Environmental characterization based on probe and reference genotypes. In: Cooper M, Hammer GL (eds) Plant adaptation and crop improvement. CAB International, Wallingford, pp 529–547
 34. Rasmusson DC (1996) Germplasm is paramount. In: Reynolds MP, Rajaram S, McNab A (eds) Increasing yield potential in wheat: breaking the barriers. CIMMYT, Mexico, pp 28–37
 35. Hayes PM, Liu BH, Knapp SJ, Chen F, Jones B, Blake T, Franckowiak J, Rasmusson D, Sorrells M, Ullrich SE, Wesenberg D, Kleinhofs A (1993) Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm. *Theor Appl Genet* 87:392–401
 36. Comadran J, Thomas WTB, van Eeuwijk FA, Ceccarelli S, Grando S, Stanca AM, Pecchioni N, Akar T, Al-Yassin A, Benbelkacem A, Ouabbou H, Bort J, Romagosa I, Hackett CA, Russell JR (2009) Patterns of genetic diversity and linkage disequilibrium in a highly structured *Hordeum vulgare* association mapping population for the Mediterranean basin. *Theor Appl Genet* 119:175–187
 37. Cavanagh C, Morell M, Mackay I, Powell P (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr Opin Plant Biol* 11:215–221
 38. Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
 39. Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539–551
 40. Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA (2011) Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proc Natl Acad Sci* 108:4488–4493
 41. Araus JL, Bort J, Steduto P, Villegas D, Royo C (2003) Breeding cereals for Mediterranean conditions: ecophysiological clues for biotechnology application. *Ann Appl Biol* 142:129–141
 42. Araus JL, Slafer GA, Royo C, Serret MD (2008) Breeding for yield potential and stress adaptation in cereals. *Crit Rev Plant Sci* 27:377–412
 43. Araus JL, Slafer GA, Reynolds MO, Royo C (2009) Breeding for quantitative variables. Part 5: breeding for yield potential. In: Ceccarelli S, Guimaraes EP, Weltzien E (eds) Participatory plant breeding. FAO, Rome, pp 449–477
 44. Cattivelli L, Ceccarelli S, Romagosa I, Stanca M (2011) Abiotic stresses in barley: problems and solutions. In: Ullrich SE (ed) Barley: improvement, production, and uses. Wiley-Blackwell, Harrisingburg, pp 282–306
 45. Sadras VO, Calderini D (2009) Crop physiology. Applications for genetic improvement and agronomy. Academic, Burlington

46. Loss SP, Siddique KHM (1994) Morphological and physiological traits associated with wheat yield increases in Mediterranean environments. *Adv Agron* 52:229–276
47. Passioura JB (2002) Environmental biology and crop improvement. *Funct Plant Biol* 29:537–546
48. Richards RA (1991) Crop improvement for temperate Australia, future opportunities. *Field Crop Res* 26:141–169
49. Worland AJ (1996) The influence of flowering time genes on environmental adaptability in European wheats. *Euphytica* 89:49–57
50. Slafer GA, Rawson HM (1994) Sensitivity of wheat phasic development to major environmental factors: a re-examination of some assumptions made by physiologists and modellers. *Aust J Plant Physiol* 21:393–426
51. Goldringer I, Prouin C, Rousset M, Galic N, Bonnin I (2006) Rapid differentiation of experimental populations of wheat for heading time in response to local climatic conditions. *Ann Bot* 98:805–817
52. Young KJ, Elliott GA (1994) An evaluation of barley accessions for adaptation to the cereal growing regions of western Australia, based on time to ear emergence. *Aust J Agr Res* 45:75–92
53. Hoogendoorn J (1985) A reciprocal F_1 monosomic analysis of the genetic control of time of ear emergence, number of leaves and number of spikelets in wheat (*Triticum aestivum* L.). *Euphytica* 34:545–558
54. Lasa JM, Igartua E, Ciudad FJ, Codesal P, Garcia EV, Gracia MP, Medina B, Romagosa I, Molina-Cano JL, Montoya JL (2001) Morphological and agronomical diversity patterns in the Spanish barley core collection. *Hereditas* 135:217–225
55. van Oosterom EJ, Acevedo E (1992) Adaptation of barley (*Hordeum vulgare* L.) to harsh Mediterranean environments. *Euphytica* 62:15–27
56. Álvaro F, Isidro J, Villegas D, García del Moral LF, Royo C (2008) Breeding effects on grain filling, biomass partitioning, and remobilization in Mediterranean durum wheat. *Agron J* 100:361–370
57. Jackson PA, Byth DE, Fischer KS, Johnston RP (1994) Genotype \times environment interactions in progeny from a barley cross: II. Variation in grain yield, yield components and dry matter production among lines with similar times to anthesis. *Field Crop Res* 37:11–23
58. Van Oosterom EJ, Kleijn DM, Ceccarelli S, Nachit MM (1993) Genotype-by-environment interactions of Barley in the Mediterranean region. *Crop Sci* 33:669–674
59. Appendino ML, Slafer GA (2003) Earliness per se and its dependence upon temperature in diploid wheat lines differing in the major gene *Eps-A^m1* alleles. *J Agric Sci* 141:149–154
60. Bullrich L, Appendino ML, Tranquilli G, Lewis S, Dubcovsky J (2002) Mapping of a thermo-sensitive earliness per se gene on *Triticum monococcum* chromosome 1A^m. *Theor Appl Genet* 105:585–593
61. Lewis S, Faricelli ME, Appendino ML, Valárik M, Dubcovsky J (2008) The chromosome region including the earliness per se locus *Eps-A^m1* affects the duration of early developmental phases and spikelet number in diploid wheat. *J Exp Bot* 59:3595–3607
62. Slafer GA (1996) Differences in phasic development rate amongst wheat cultivars independent of responses to photoperiod and vernalization. A viewpoint of the intrinsic earliness hypothesis. *J Agric Sci* 126:403–419
63. Slafer GA, Rawson HM (1995) Base and optimum temperatures vary with genotype and stage of development in wheat. *Plant Cell Environ* 18:671–679
64. Cockram J, Jones H, Leigh FJ, O'Sullivan D, Powell W, Laurie DA, Greenland A (2007) Control of flowering time in temperate cereals, genes, domestication, and sustainable productivity. *J Exp Bot* 58:1231–1244
65. Eagles HA, Cane K, Vallance N (2009) The flow of alleles of important photoperiod basically and vernalisation genes through Australian wheat. *Crop Pasture Sci* 60:646–657
66. Baum M, Grando S, Backes G, Jahoor A, Sabbagh A, Ceccarelli S (2003) QTLs for agronomic traits in the Mediterranean environment identified in recombinant inbred lines of the cross 'Arta' \times *H. spontaneum* 41-1. *Theor Appl Genet* 107:1215–1225
67. Bezant J, Laurie D, Pratchett N, Chojecski J, Kearsey M (1996) Marker regression mapping of QTL controlling flowering time and plant height in a spring barley (*Hordeum vulgare* L.) cross. *Heredity* 77:64–73
68. Li JZ, Huang XQ, Heinrichs F, Ganai MW, Röder MS (2006) Analysis of QTLs for yield components, agronomic traits and disease resistance in an advanced backcross population of spring barley. *Genome* 49:454–466
69. Tinker NA, Mather DE, Blake TK, Briggs KG, Choo TM, Dahleen L, Dofing SM, Falk DE, Ferguson T, Frankowiak JD, Graf R, Hayes PM, Hoffman D, Irvine RB, Kleinhofs A, Legge W, Rosnagel BG, Saghai Maroof MA, Scoles GJ, Shugar LP, Steffenson B, Ullrich S, Kasha KJ (1996) Regions of the genome that affect agronomic performance in two-row barley. *Crop Sci* 36:1053–1062
70. Cuesta-Marcos A, Casas AM, Hayes PM, Gracia MP, Lasa JM, Ciudad F, Codesal P, Molina-Cano JL, Igartua E (2009) Yield QTL affected by heading date in Mediterranean grown barley. *Plant Breeding* 128:46–53
71. Romagosa I, Ullrich SE, Han F, Hayes PM (1996) Use of the AMMI model in QTL mapping for adaptation in barley. *Theor Appl Genet* 93:30–37
72. Malosetti M, Voltas J, Romagosa I, Ullrich SE, van Eeuwijk FA (2004) Mixed models including environmental variables for studying QTL by environment interaction. *Euphytica* 137:139–145
73. Beales J, Turner A, Griffiths S, Snape JW, Laurie DA (2007) A *Pseudo-Response Regulator* is misexpressed in the photoperiod insensitive *Ppd-D1a* mutant of wheat (*Triticum aestivum* L.). *Theor Appl Genet* 115:721–733
74. Wilhelm EP, Turner AS, Laurie DA (2009) Photoperiod insensitive *Ppd-A1a* mutations in tetraploid wheat (*Triticum durum* Desf.). *Theor Appl Genet* 118:285–294
75. Laurie DA, Pratchett N, Bezant JH, Snape JW (1995) RFLP mapping of five major genes and eight quantitative trait loci controlling flowering time in a winter \times spring barley (*Hordeum vulgare* L.) cross. *Genome* 38:575–585

76. Jones H, Leigh FJ, Mackay I, Bower MA, Smith LMJ, Charles MP, Jones G, Jones MJ, Brown TA, Powell W (2008) Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated East of the Fertile Crescent. *Mol Biol Evol* 25:2211–2219
77. Faure S, Higgins J, Turner A, Laurie DA (2007) The flowering locus T-like gene family in barley (*Hordeum vulgare*). *Genetics* 176:599–609
78. Fu D, Szűcs P, Liuling Y, Helguera M, Skinner JS, Zitzewitz J, Hayes PM, Dubcovsky J (2005) Large deletions within the first intron in *Vrn-1* are associated with spring growth habit in barley and wheat. *Mol Genet Genomics* 273:54–65
79. Trevaskis B, Bagnall DJ, Ellis MH, Peacock WJ, Dennis ES (2003) MADS box genes control vernalization-induced flowering in cereals. *Proc Natl Acad Sci* 22:13099–13104
80. Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *Vrn1*. *Proc Natl Acad Sci* 100:6263–6268
81. Distelfeld A, Tranquilli G, Chengxia L, Yan L, Dubcovsky J (2009) Genetic and molecular characterization of the *Vrn2* loci in tetraploid wheat. *Plant Physiol* 149:245–257
82. Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, San Miguel P, Bennetzen JL, Echenique V, Dubcovsky J (2004) The wheat *Vrn2* gene, a flowering repressor down-regulated by vernalization. *Science* 303:1640–1644
83. Casao MC, Igartua E, Karsai I, Bhat PR, Cuadrado N, Gracia MP, Lasa JM, Casas AM (2011) Introgression of an intermediate VRNH1 allele in barley (*Hordeum vulgare* L.) leads to reduced vernalization requirement without affecting freezing tolerance. *Mol Breed*. doi:10.1007/s11032-010-9497
84. Bonnin I, Rousset M, Madur D, Sourdille P, Dupuits C, Brunel D, Goldringer I (2008) FT genome A and D polymorphisms are associated with the variation of earliness components in hexaploid wheat. *Theor Appl Genet* 116:383–394
85. Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene *Vrn3* is an orthologue of *FT*. *Proc Natl Acad Sci* 103:19581–19586
86. Börner A, Buck-Sorlin GH, Hayes PM, Malyshev S, Korzun V (2002) Molecular mapping of major genes and quantitative trait loci determining flowering time in response to photoperiod in barley. *Plant Breed* 121:129–132
87. Franckowiak JD (1997) Revised linkage maps for morphological markers in barley, *Hordeum vulgare*. *Barley Genet Newsl* 26:9–21
88. Lundqvist U, Franckowiak JD, Konishi T (1997) New and revised descriptions of barley genes. *Barley Genet Newsl* 26:22–516
89. Stracke S, Börner A (1998) Molecular mapping of the photoperiod response gene *ea7* in barley. *Theor Appl Genet* 97:797–800
90. Chen A, Baumann U, Fincher GB, Collins NC (2009) *Flt-2L*, a locus in barley controlling flowering time, spike density, and plant height. *Funct Integr Genomics* 9:243–254
91. Scarth R, Law CN (1983) The location of the photoperiodic gene, *Ppd2*, and an additional factor for ear-emergence time on chromosome 2B of wheat. *Heredity* 51:607–619
92. Shindo C, Tsujimoto H, Sasakuma T (2003) Segregation analysis of heading traits in hexaploid wheat utilizing recombinant inbred lines. *Heredity* 90:56–63
93. Yoshida T, Nishida H, Zhu J, Nitcher R, Distelfeld A, Akashi Y, Kato K, Dubcovsky J (2010) *Vrn-D4* is a vernalization gene located on the centromeric region of chromosome 5D in hexaploid wheat. *Theor Appl Genet* 120:543–552
94. Kato K, Miura H, Sawada S (2002) Characterization of *QEet.ocs-5A.1*, a quantitative trait locus for ear emergence time on wheat chromosome 5AL. *Plant Breed* 121:389–393
95. Law CN, Worland AJ (1997) Genetic analysis of some flowering time and adaptive traits in wheat. *New Phytol* 137:19–28
96. Griffiths S, Simmonds J, Leverington M, Wang Y, Fish L, Sayers L, Alibert L, Orford S, Wingen L, Herry L, Faure S, Laurie D, Bilham L, Snape J (2009) Meta-QTL analysis of the genetic control of ear emergence in elite European winter wheat germplasm. *Theor Appl Genet* 119:383–395
97. Kuchel H, Hollamby G, Langridge P, Williams K, Jefferies SP (2006) Identification of genetic loci associated with ear-emergence in bread wheat. *Theor Appl Genet* 113:1103–1112
98. Sourdille P, Snape JW, Charmet G, Nakata N, Bernard S, Bernard M (2000) Detection of QTLs for heading time and photoperiod response in wheat using a doubled-haploid population. *Genome* 43:487–494
99. Boer M, Wright D, Feng L, Podlich D, Luo L, Cooper M, van Eeuwijk FA (2007) A mixed model QTL analysis for multiple environment trial data using environmental covariables for QTLxE, with an example in maize. *Genetics* 177:1801–1813
100. Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM (2010) *GenStat for windows* (13th edition) introduction. VSN International, Hemel Hempstead
101. Limin A, Corey A, Hayes PM, Fowler DB (2007) Low-temperature acclimation of barley cultivars used as parents in mapping populations: response to photoperiod, vernalization and phenological development. *Planta* 226:139–146
102. Slafer GA, Abeledo LG, Miralles DJ, González FG, Whitechurch EM (2001) Photoperiod sensitivity during stem elongation as an avenue to raise potential yield in wheat. *Euphytica* 119:191–197
103. Slafer GA, Araus JL, Royo C, García del Moral LF (2005) Promising eco-physiological traits for genetic improvement of cereal yields in Mediterranean environments. *Ann Appl Biol* 146:61–70
104. Fischer RA (2007) Understanding the physiological basis of yield potential in wheat. *J Agric Sci* 145:99–113
105. Miralles DJ, Slafer GA (2007) Sink limitations to yield in wheat, how could it be reduced? *J Agric Sci* 145:139–149
106. Halloran GM, Pennell AL (1982) Duration and rate of development phases in wheat in two environments. *Ann Bot* 49:115–121
107. Whitechurch EM, Slafer GA, Miralles DJ (2007) Variability in the duration of stem elongation in wheat and barley genotypes. *J Agron Crop Sci* 193:138–145
108. Borràs-Gelonch G, Rebetzke G, Richards R, Romagosa I (2011b) Genetic control of duration of pre-anthesis phases in wheat (*Triticum aestivum* L.) and relationships to leaf appearance, tillering and dry matter accumulation. *Journal of Experimental Botany*, doi: 10.1093/jxb/err230

109. Appleyard M, Kirby EJM, Fellowes G (1982) Relationships between the duration of phases in the pre-anthesis life cycle of spring barley. *Aust J Agr Res* 33:917–925
110. Borràs G, Romagosa I, van Eeuwijk F, Slafer GA (2009) Genetic variability in the duration of pre-heading phases and relationships with leaf appearance and tillering dynamics in a barley population. *Field Crop Res* 113:95–104
111. Borràs-Gelonch G, Slafer GA, Casas A, van Eeuwijk F, Romagosa I (2010) Genetic control of pre-heading phases and other traits related to development in a double haploid barley population (*Hordeum vulgare* L.). *Field Crop Res* 119:36–47
112. Kernich GC, Halloran GM, Flood RG (1995) Variation in development patterns of wild barley (*Hordeum spontaneum* L) and cultivated barley (*H vulgare* L). *Euphytica* 82:105–115
113. Kernich GC, Halloran GM, Flood RG (1997) Variation in duration of pre-anthesis phases of development in barley (*Hordeum vulgare*). *Aust J Agr Res* 48:59–66
114. Kitchen BM, Rasmusson DC (1983) Duration and inheritance of leaf initiation, spike initiation and spike growth in barley. *Crop Sci* 23:939–943
115. González FG, Slafer GA, Miralles DJ (2002) Vernalization and photoperiod responses in wheat pre-flowering reproductive phases. *Field Crop Res* 74:183–195
116. Miralles DJ, Richards RA (2000) Responses of leaf and tiller emergence and primordium initiation in wheat and barley to interchanged photoperiod. *Ann Bot* 85:655–663
117. González FG, Slafer GA, Miralles DJ (2005) Pre-anthesis development and number of fertile florets in wheat as affected by photoperiod sensitivity genes *Ppd-D1* and *Ppd-B1*. *Euphytica* 146:253–269
118. Zhou Y, Li W, Wu W, Chen Q, Mao D, Worland AJ (2001) Genetic dissection of heading time and its components in rice. *Theor Appl Genet* 102:1236–1242
119. Borràs-Gelonch G, Denti M, Thomas WTB, Romagosa I (2011a) Genetic control of pre-heading phases in the Steptoe x Morex barley population under different conditions of photoperiod and temperature. *Euphytica* doi: 10.1007/s10681-011-0526-7
120. Denis JB (1988) Two-way analysis using covariates. *Statistics* 19:123–132
121. van Eeuwijk FA, Denis JB, Kang MS (1996) Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In: Kang MS, Gauch HG (eds) *Genotype-by-environment interaction*. CRC Press, Boca Raton, pp 15–50
122. Yates F, Cochran WG (1938) The analysis of groups of experiments. *J Agric Sci* 28:556–580
123. Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant breeding programme. *Aust J Agr Res* 14:742–754
124. van Eeuwijk FA (1995) Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. *Euphytica* 84:1–7
125. Denis JB, Gower JC (1996) Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. *Appl Stat* 45:479–492
126. Gabriel KR (1998) Generalised bilinear regression. *Biometrika* 85:689–700
127. van Eeuwijk FA (1995) Multiplicative interaction in generalized linear models. *Biometrics* 51:1017–1032
128. Gabriel KR (1978) Least squares approximation of matrices by additive and multiplicative models. *J Roy Stat Soc Ser B* 40:186–196
129. Gauch HG (1988) Model selection and validation for yield trials with interaction. *Biometrics* 44:705–715
130. Gollob HF (1968) A statistical model which combines features of factor analysis and analysis of variance techniques. *Psychometrika* 33:73–115
131. Mandel J (1969) The partitioning of interaction in analysis of variance. *J Res NBS* 73B:309–328
132. Yan W, Kang MS (2003) *GGE biplot analysis: a graphical tool for breeders, geneticists, and agronomists*. CRC Press, Boca Raton
133. Yan W, Hunt LA, Sheng Q, Szlavnics Z (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci* 40:597–605
134. Cornelius PL (1993) Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. *Crop Sci* 33:1186–1193
135. Kempton RA (1984) The use of biplots in interpreting variety by environment interactions. *J Agric Sci* 103:123–135
136. Molina-Cano JL, García del Moral LF, Ramos JM, García del Moral MB, Romagosa I, Roca de Togores F (1990) Quantitative phenotypic expression of three mutant genes in barley and the basis for defining an ideotype for Mediterranean environments. *Theor Appl Genet* 80:762–768
137. Romagosa I, Fox PN, del Moral G, Ramos JM, García del Moral B, Roca de Togores F, Molina-Cano JL (1993) Integration of statistical and physiological analyses of adaptation of near-isogenic barley lines. *Theor Appl Genet* 86:822–826
138. Vargas M, Crossa J, van Eeuwijk FA, Ramírez ME, Sayre K (1999) Using AMMI, factorial regression, and partial least squares regression models for interpreting genotype × environment interaction. *Crop Sci* 39:955–967
139. Voltas J, van Eeuwijk FA, Sombrero A, Lafarga A, Igartua E, Romagosa I (1999) Integrating statistical and ecophysiological analysis of genotype by environment interaction for grain filling of barley in Mediterranean areas I. Individual grain weight. *Field Crop Res* 62:63–74
140. Voltas J, van Eeuwijk FA, Araus JL, Romagosa I (1999) Integrating statistical and ecophysiological analysis of genotype by environment interaction for grain filling of barley in Mediterranean areas II. Grain growth. *Field Crop Res* 62:75–84
141. Royo C, Rodríguez A, Romagosa I (1993) Differential adaptation of complete and substituted triticale to acid soils. *Plant Breed* 111:113–119
142. Lynch M, Walsh JB (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland
143. van Eeuwijk FA, Crossa J, Vargas M, Ribaut JM (2001) Variants of factorial regression for analysing QTL by environment interaction. In: Gallais A, Dillmann C, Goldringer I (eds) *Eucarpia, quantitative genetics and breeding methods: the*

way ahead, vol 96, INRA Editions Versailles Les Colloques series. INRA, Paris, pp 107–116

144. van Eeuwijk FA, Crossa J, Vargas M, Ribaut JM (2002) Analysing QTL by environment interaction by factorial regression, with an application to the CIMMYT drought and low nitrogen stress programme in maize. In: Kang MS (ed) Quantitative genetics, genomics and plant breeding. CAB International, Wallingford, pp 245–256
145. Piepho HP, Pillen K (2004) Mixed modeling for QTL \times environment interaction analysis. *Euphytica* 137:147–153
146. IRRI (2008) CropStat for windows, version 5. Biometrics and Bioinformatics Unit, International Rice Research Institute, Los Baños, Philippines. <http://www.irri.org/science/software/irristat.asp>. Accessed 27 Feb 2008
147. Gauch HG (2007) MATMODEL version 3.0: Open source software for AMMI and related analyses. Crop and Soil Sciences, Cornell University, Ithaca. <http://www.css.cornell.edu/staff/gauch>. Accessed 27 Feb 2008
148. Balzarini MG, Bruno C, Peña A, Teich I, Di Rienzo JA (2010) Estadística en Biotecnología. Aplicaciones en InfoGen. Grupo InfoStat, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Argentina
149. Di Rienzo JA, Casanoves F, Balzarini MG, Gonzalez L, Tablada M, Robledo CW (2008) InfoStat, versión 2008. Grupo InfoStat, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Argentina
150. Kang MS (2003) Handbook of formulas and software for plant geneticists and breeders. Haworth Press, Binghamton
151. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9:432–441
152. Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090
153. Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
154. Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 9:166–177
155. Bertin N, Martre P, Génard M, Quilot B, Salon C (2010) Under what circumstances can process-based simulation models link genotype to phenotype for complex traits? Case-study of fruit and grain quality traits. *J Exp Bot* 61:955–967
156. Chenu K, Chapman SC, Tardieu F, McLean G, Welcker C, Hammer GL (2009) Simulating the yield impacts of organ-level quantitative trait loci associated with drought response in maize – a ‘gene-to-phenotype’ modeling approach. *Genetics* 183:1507–1523
157. Hammer G, Cooper M, Tardieu F, Welch S, Walsh B, van Eeuwijk FA, Chapman S, Podlich D (2006) Models for navigating biological complexity in breeding improved crop plants. *Trends Plant Sci* 11:587–593

Books and Reviews

- Ceccarelli S, Guimaraes EP, Weltzien E (eds) (2009) Participatory plant breeding. FAO, Rome

Geochemical Modeling in Environmental and Geological Studies

CHEN ZHU

Department of Geological Sciences, Indiana University, Bloomington, IN, USA

Article Outline

Glossary
 Definition of the Subject and Its Importance
 Introduction
 What Is Geochemical Modeling?
 The Part of Three
 Future Directions
 Acknowledgment
 Bibliography

Glossary

Activation energy The energy that must be overcome in order for a chemical reaction to occur.

Mass transport The net movement of mass from one location to another due to hydrological processes such as advection, dynamic dispersion, chemical reactions, and microbial activities.

Rate law A rate law is a statement about how the rate of a reaction depends on the concentrations of the participating species.

Solubility product Equilibrium constants for various kinds of reactions with a solid phase on one side and its constituent ions on the other.

Species A chemical entity distinguishable from other entities by molecular formula and structure, e.g., CO_2 and O_2 in a gas, and HCO_3^- , $\text{H}_2\text{CO}_3^0_{(\text{aq})}$, CO_3^{2-} , $\text{NaHCO}_3^0_{(\text{aq})}$.

Definition of the Subject and Its Importance

Geochemical modeling uses a set of mathematical expressions thought to represent chemical and transport processes in a particular geological system. The predictions of the model are partially observable or experimentally verifiable. Geochemical modeling has found applications in studies of chemical reactions in geological and

environmental systems because of its utilities for synthesis of data, testing scenarios, and predicting long-term consequences of chemical reactions.

Introduction

Geochemical modeling is a powerful and indispensable tool for research and investigations of environmental sustainability science and technology. It allows quantitative evaluation of complex processes that often have feedback loops and it can predict the extent and consequences of geochemical reactions in the order of thousands to tens of thousands of years, beyond the range of laboratory experiments.

An excellent example of the utility of geochemical modeling in environmental sustainability science and engineering and its unique characteristics is illustrated by geological carbon sequestration – the injection of carbon dioxide (CO_2) into deep geological formations as a climate mitigation tool. Upon injection of CO_2 into a geological formation, CO_2 is dissolved into the native brine. The carbonated brine becomes acidic and corrosive, aggressively reacting with host rocks (e.g., a sandstone or carbonate rocks). Some primary (native) minerals are dissolved and secondary minerals precipitated. These dissolution-precipitation reactions can drastically change the porosity and permeability of the host rocks, and thereby impact the injectivity and storage safety in the long-term.

In the above example, there are many processes that are coupled. The flow of the separate supercritical CO_2 phase transports the CO_2 in the aquifer, determined by the viscosity, density, and the relative permeability of CO_2 in contrast to the brine. CO_2 dissolves into the brine when CO_2 makes contact with it. A brine with dissolved CO_2 , probably in the form of HCO_3^- , has higher density than the rest. Density-driven vertical convection can occur, and convection brings about chemical gradients that result in more reactions.

The above example demonstrates that the system is complex and that the chemical reactions are coupled with each other and also coupled to transport processes, such as advection and dispersion. It is difficult to quantitatively evaluate these reactions without a computer model. There may be some ideas about what reactions will occur and how fast they will occur, but by developing a model, ideas can be

formulated explicitly and thoughts can be tested quantitatively. Often, ideas are restricted to a particular aspect of chemical reactions, gleaned from specific laboratory experiments or field observations. Whether such ideas are valid when they are examined with measurable or observable consequences of the overall chemical systems is not known. In other words, whether the ideas of a subsystem hold up in the overall scheme of things must be tested. The system of concern often not only involves the coupling of hundreds of reactions, but also to processes like diffusion, advection, and dispersion, biological activities, thermal conduction, mechanical stress, and deformation.

The geological carbon sequestration example also illustrates the necessity of the prediction of chemical reactions for all practical purposes. Although federal or national regulations for safe underground CO_2 storage still need to emerge, it is reasonable to assume that national and local regulations will demand risk assessments of wellbore integrity, well injectivity, and long-term performance in the order of thousands of years. The geology of each injection site differs, and geological heterogeneities at a given site are a fact of life. It is not possible to conduct laboratory experiments either for each possible geochemical system or for durations of more than a few months. Performance assessments necessary at all stages of CO_2 storage operations (site assessment/selection, design, installation, operations and monitoring, and closure/post-closure) have to be partly based on geochemical modeling predictions.

Thus, it is clear that geochemical modeling is an indispensable tool in geochemical research and engineering. This entry introduces the basic concepts of geochemical modeling, provides information for accessing modeling codes and further readings, and shows applications in the field of environmental sustainability sciences and technology.

What Is Geochemical Modeling?

To modify Zhu and Anderson's [1] definition a little bit, a *geochemical model* is "an abstract object, described by a set of mathematical expressions thought to represent chemical and transport processes in a particular system. The predictions of the model are partially observable or experimentally verifiable."

A geochemical model typically includes a *geochemical reaction network*, which means the finite array of reactions in a geochemical system and transport processes for a reactive transport system.

Mathematically, for a geochemical system that has n species, the following ordinary differential equations completely define the geochemical reaction network [2],

$$\frac{dm_i}{dt} = \sum_j v_{ij} r_{ij}, i \in n, \quad (1)$$

where m_i denotes the concentrations of i th species, t the time, v_j the stoichiometric coefficient for i th species in the j th reaction, and r_{ij} the production or consumption rate of the i th species in the j th reaction. For a reactive transport system, the geochemical reaction network is defined by the transport equations,

$$\frac{\partial m_i}{\partial t} + L(C_i) = \sum_j v_j r_i \quad (2)$$

where L is the advection, dispersion, diffusion operator [3].

For historic development of geochemical modeling, the readers are referred to Zhu and Anderson [1] and Nordstrom [4]. To develop or apply a geochemical model, the modeler needs three parts: (1) a computer code that solves Eqs. 1 and/or 2; (2) a thermodynamic and kinetics database; (3) an input file that supplies the chemical analysis and the design or conceptual model of the geochemical model. Zhu and Anderson [1] called it “the part of three.”

The Part of Three

Numerous computer programs for geochemical modeling have been developed. Without exception, the Newton–Raphson iteration method is used to solve the highly nonlinear Eq. 1. These computer codes include EQ3/6 [5], SOLMINEQ.88 [6], PHREEQC [7], and MINTEQA2 [8]. The codes that are sponsored by government agencies are essentially free of charge and are widely used. With ever more increasing computing power, tremendous advancements have been seen in code developments in the last decade. Zhu [9] pointed out that computer code developments are ahead of underlying science.

Most widely distributed computer codes come with a database of equilibrium constants for chemical reactions. For example, a database called DATA0.DAT was initially developed for EQ3/6 by Tom Wolery [5] and Jim Johnson [10, 11]. This database was later adopted for the program PHREEQC as LLNL.DAT, GWB as THERMO.DAT, and the database in ToughReact. Similarly, thermodynamic databases have been developed for MINTEQA2 as MINTEQA2.DAT for PHREEQC as PHREEQC.DAT.

The compilations of equilibrium constants for chemical reactions draw from the available databases of standard state properties for minerals, such as those from Helgeson et al. [12], Wagman et al. [13], Berman [14, 15], Holland and Powell [16], Nordstrom et al. [17], and Robie and Hemingway [18]. More specialized databases are available for uranium [19]. Because these internally consistent databases only contain a limited number of minerals, while applications of geochemical modeling to a variety of geological and environmental topics require a wider range of minerals and solids, additional minerals are added to these equilibrium constant databases MINTEQA2 and PHREEQC.

For aqueous species, an internally consistent database developed by Harold Helgeson and collaborators, which includes a large number of aqueous species, has been widely used in the field of geochemistry [20–23]. In this database, the temperature and pressure dependences of thermodynamic properties for aqueous species were predicted using the parameters of the revised Helgeson–Kirkham–Flowers (HKF) equations of state for aqueous species [20, 24, 25]. Activity coefficients for the charged aqueous species were calculated from the extended Debye–Hückel equation or B-dot equation fitted to mean salt NaCl activity coefficients [26]. The computer program SUPCRT92 can be used to generate equilibrium constants at elevated temperatures and pressures [11]. Equilibrium constants have been calculated using the standard state properties from this database and compiled into databases that accompany the program EQ3/6 which was then adopted to other programs, e.g., LLNL.DAT in PHREEQC and THERMO.DAT in GWB®.

It has been known for quite some time that the internal consistency of thermodynamic data for modeling calculations is important. Nordstrom and Munoz [27] elaborated on the topic of internal consistency and

readers are urged to consult their writing on the topic. Note that the standard state thermodynamic properties in the mineral databases [12, 14–17] are *internally consistent*, but some added minerals may not necessarily be so. The aqueous species collected in the Helgeson–Sverjensky–Shock compilations are internally consistent, but others collected may not be.

In the deep parts of the subsurface and in the shallow parts where evaporate sediment beds are located, groundwater becomes briny and can have concentrations of dissolved solids up to 300,000 mg/L. In dealing with concentrated solutions, Pitzer's ion interaction approach using virial specific interaction equations is generally preferred over the ion association theory when calculating ionic activities. The Pitzer's model, commonly the Harvie–Moller–Weare (HMW) formulation of it [28], has been incorporated into geochemical modeling codes EQ3/6, PHRQPITZ [29], and TOUGHREACT [30]. Although progress has been made in compiling the Pitzer interaction parameters [31], the lack of Pitzer's activity coefficient parameters at elevated temperatures and for minor or trace elements remains a barrier to accurate calculation of solubility and saturation indices for highly saline fluids.

The input file is where the modeler defines the composition of the chemical system, includes or excludes certain types of chemical reactions, and assigns the boundary and initial conditions. In other words, this is where the modeler translates the conceptual model into recognizable formats by computer programs. The modeler also develops the conceptual model in this way [1].

From the discussion of each of the “part of three,” it should be clear that a computer program should always be distinguished from a geochemical model. For example, PHREEQC is a geochemical modeling computer program or code; it is not a geochemical model. It is always true that the modeler, not the computer code, produces a geochemical model. A geochemical modeling code is a utility, e.g., an oven. The oven does not bake the cake. The chef bakes the cake.

Speciation–Solubility Modeling

Speciation modeling calculates the distribution of aqueous species and mineral saturation indices according to the solutions of the mass and charge

balance and mass action equations. From the activities of the aqueous species, one can calculate Saturation Indices (*SI*),

$$SI = \log \left(\frac{IAP}{K} \right) \quad (3)$$

where *K* stands for the equilibrium constant of the dissolution reaction and *IAP* stands for the Ion Activity Product. When *SI* = 0, the mineral is at equilibrium with the aqueous solution. When *SI* < 0, the aqueous solution is undersaturated with respect to the mineral of concern and the mineral will dissolve. When *SI* > 0, the aqueous solution is supersaturated with respect to the mineral and the mineral will precipitate.

Speciation–solubility modeling has become a routine exercise since Garrels and Thompson [32] first calculated the aqueous speciation in seawater and saturation states with respect to mineral solubility. The principles are well known, and the numerical modeling techniques and their tweaking are mature. There are hundreds of computer codes available for this kind of calculation.

The results of speciation and solubility calculations are used for a number of issues in environmental sustainability science and technology. The toxicity of many chemicals is not only related to the total concentrations, but also to the species. Liu et al. [33] measured the total concentrations of antimony (Sb) with valence of V. They then used speciation calculations and found the dominant aqueous Sb species was $\text{Sb}(\text{OH})_6^-$. In general, it is assumed that aqueous species in the solution are in mutual equilibrium (homogeneous equilibria). The exception to this rule is redox species, which are well known for not being at equilibrium in surficial water bodies [34, 35].

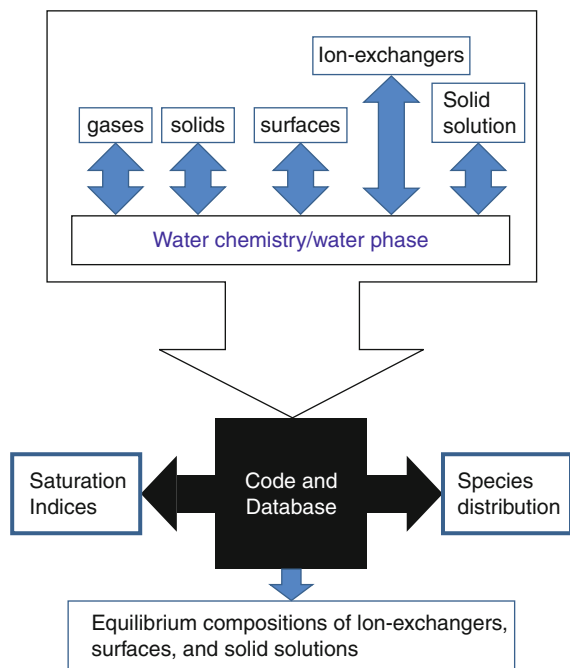
The calculated saturation indices give directions of chemical reactions. As elaborated in Zhu and Anderson [1] and in numerous books on thermodynamics, saturation indices show the direction of the chemical reactions, as dictated by the second law of thermodynamics, but tell nothing about the rate of reactions. For example, it is well known in geochemistry that natural waters are often supersaturated with respect to crystalline quartz, but the rate of quartz precipitation is too slow for these waters to reach equilibrium with quartz even after thousands of years at low temperatures.

However, reaction direction is the first thing that one must know before one can extract reaction rates (one needs to know which direction the reactions will go before one can estimate how fast or slow they will do it).

One could extend the scope of traditional speciation–solubility modeling to include the equilibrium partitioning of a chemical between the aqueous solution and mineral surfaces (Fig. 1). Many computer codes (e.g., MINTEQA2 and PHREEQC) now allow the calculations of surface adsorption according to the surface complexation theory [35–37]. One can find the details of the surface complexation theories in the textbooks cited above. In terms of modeling, when the total concentration of an ion (e.g., Pb^{2+}) in an aqueous solution is given, the codes can calculate how much Pb^{2+} is partitioned onto the mineral surface(s), how much Pb^{2+} remains in the aqueous solution, what the relative percentage of surface bound Pb^{2+} among different surface species is, and what the dominant aqueous Pb^{2+} species are.

Similarly, one can calculate the equilibrium partitioning of chemicals between an aqueous solution and ion-exchangers. Unlike the surface adsorption and the surface complexation theory that describes it, modeling of ion-exchange reactions lacks both theoretical footing and internally consistent databases [1]. Readers are encouraged to read more in Appelo and Postma [38]. Equally, one can also calculate the equilibrium distribution between an aqueous solution and solid solution phases [49, 50], and gas phases (Fig. 1).

Speciation–solubility modeling provides a “snapshot” of a dynamic system, and the basic building block for more advanced process modeling. The calculated activities of the various ionic and molecular species give the *IAP* for the saturation state evaluation. This type of calculation represents the majority of applications of geochemical modeling to the field of environmental sustainability science and technology. However, despite the maturity of the modeling techniques, a number of challenges exist which make the evaluation of the reaction directions a nontrivial task (cf. [9]).



Geochemical Modeling in Environmental and Geological Studies. Figure 1

Schematic configuration of speciation–solubility models

Reaction Path Modeling

Once the results of aqueous speciation and solubility calculations are obtained (i.e., calculated species distribution and saturation indices) for a given temperature, pressure, and instance of time, processes can be modeled. The simplest next step is *reaction path modeling*, tracing the evolution of aqueous solution composition and speciation and mineral paragenesis through time or the reaction progress as a result of irreversible reactions (e.g., feldspar dissolution) or processes (e.g., titration, mixing, or increase or decrease of temperature or pressure). The modeling is accomplished by applying the principle of mass balance, thermodynamics that govern the equilibrium between species, and kinetics that govern the rate of mass transfer among phases. The concept and mathematical foundation of reaction path modeling was introduced to geochemistry by Harold Helgeson [39]. Numerous articles and books have described this approach [2, 5, 40, 41]. Computer codes EQ3/6, PHREEQC, MINTEQA2, SOLMINEQ.88 and GWB[®] can all perform these kinds of calculations.

In the past a few years, advancement has been seen in computing powers and development of computer programs, which make it possible to perform reaction path modeling involving complicated reaction networks. This includes incorporation of various forms of rate laws and inclusion of an almost unlimited number of reactions into a single model. Conceptual developments in geochemical reaction networks now allow one to explore the intricacies of feedback mechanisms for complex inorganic geochemical systems [42].

One area that has been rapidly advanced in recent years is the linking of microbial activities with a network of redox and non-redox reactions and the exploration of the complex feedback loops in biogeochemistry. For example, Istok et al. [43] carried out a *biogeochemical reaction path modeling* for simulating an in situ field experiment which investigated the bioreduction of uranium near Oak Ridge National Laboratory in Tennessee, USA. In the field experiment, ethanol was injected into the aquifer to stimulate microbial activities at the field site and the reduction of nitrate, U(VI), Fe(III), Mn(IV), and sulfate were observed to proceed concomitantly.

Historically, microbial mediated reactions were modeled as semiempirical kinetics, decoupled into a whole suit of inorganic geochemical reactions as a consequence of biostimulation or bioremediation [43, 44]. The Monod-type kinetic expressions are used to describe rates of substrate utilization and biomass production. Istok et al. [43] developed a new approach, dubbed as “thermodynamically based.” In their approach, the actual microbial community is represented by a synthetic microbial community consisting of a collection of microbial groups, each with a unique growth equation that couples a specific pair of energy yielding redox reactions. Simulations monitored temporal changes of microbial biomass, community composition, aqueous speciation, and oxidation states of multivalent chemicals, as well as the dissolution and precipitation of minerals.

Istok et al.’s [43] modeling results are shown in Fig. 2. Simulations predicted that acetate addition will result in the growth of only 8 of the 25 microbial groups during the experiment. The increasing biomass and changing community composition occurred as increasing amounts of acetate were reacted. Early on in the reaction path, predicted biomass increase was

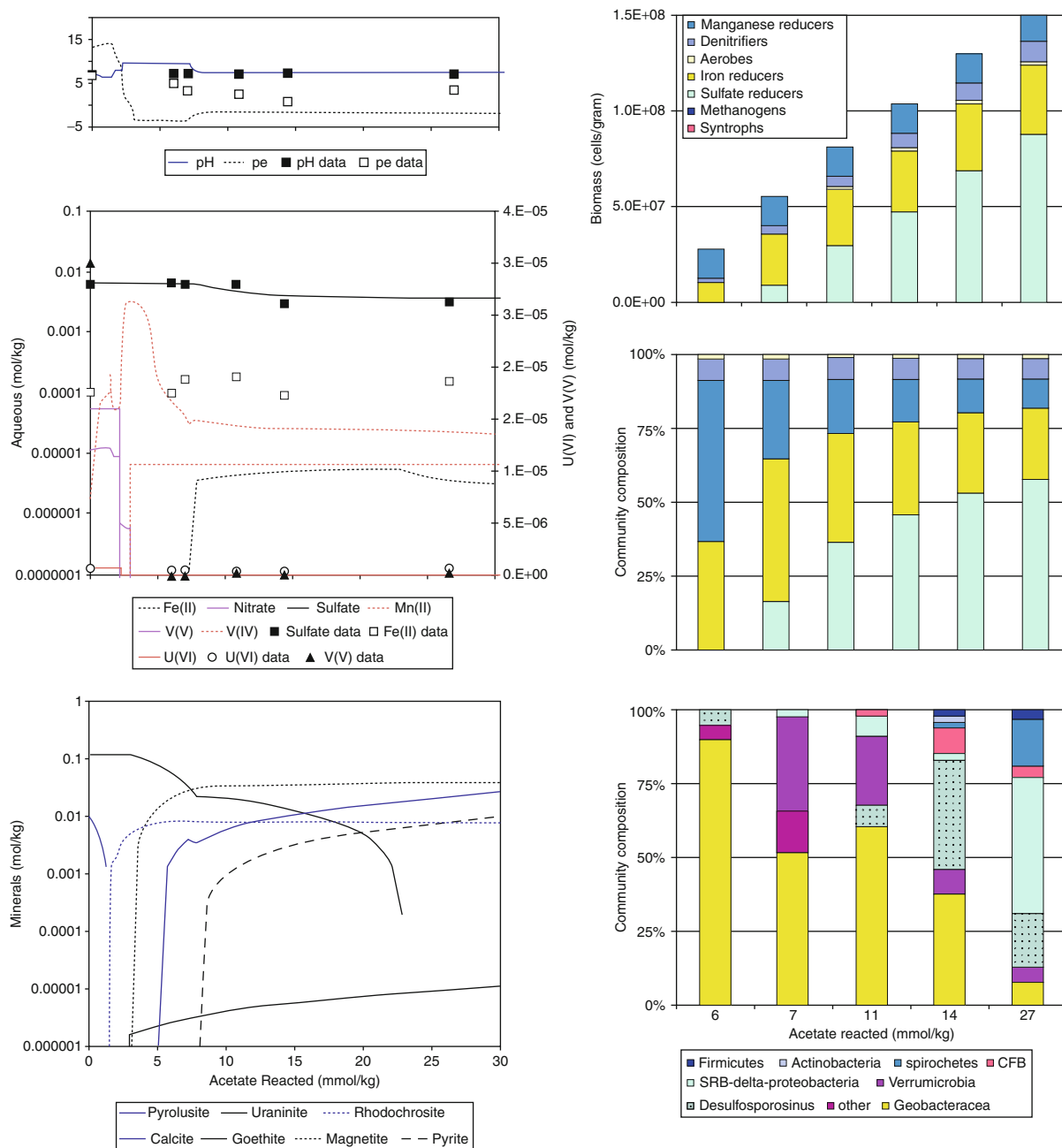
dominated by predicted growth of manganese reducers, iron reducers, and denitrifiers, with much smaller predicted growth of other groups, reflecting the relatively larger initial amounts of Mn(IV)- and Fe(III)-bearing minerals and nitrate compared to oxygen and sulfate. As additional acetate was reacted, growth of other groups was predicted to become energetically favorable, especially sulfate reducers. Predicted patterns of the growth of the various groups resulted in predicted changes in community composition. The results are generally consistent with clone libraries developed from groundwater samples.

While some computer software is fully capable of performing such complex computations, the requirements of modeling parameters for the biogeochemical reactions are daunting [45]. Numerous nonunique interpretations of field data may be possible. However, geochemical modeling can help to integrate processes and reactions in a quantitative manner and provide a more comprehensive understanding of biogeochemical processes than simply providing apparent zeroth-order and first-order rates.

Coupled Reactive Mass Transport Modeling

Geochemical models included in this category all solve the advection-dispersion-reaction (ADR) mass transport equations, typically using the sequential iteration approach. The reaction term is fully coupled to chemical equilibrium and kinetics (and microbial activities). In other words, at each time step and at each node or in each grid cell, reaction path calculations described above are performed. Many transport codes on the market are termed “reactive transport model,” but only the partitioning coefficient or K_D approach is used. In the K_D approach, all chemical reactions pertinent to a chemical are described by a single parameter [46]. In this entry, only multiple component mass transport models with the nonlinear mass balance equations for speciation and mass transfer coupled to the ADR equation are called coupled reactive mass transport (CRMT) models. For the fundamentals of the subject, readers are referred to Yeh and Tripathi [47].

Some codes in this category are coupled to fluid flow and also have the capabilities of simulating multiphase flow (air, water, and CO_2), density-dependent flow,



Geochemical Modeling in Environmental and Geological Studies. Figure 2

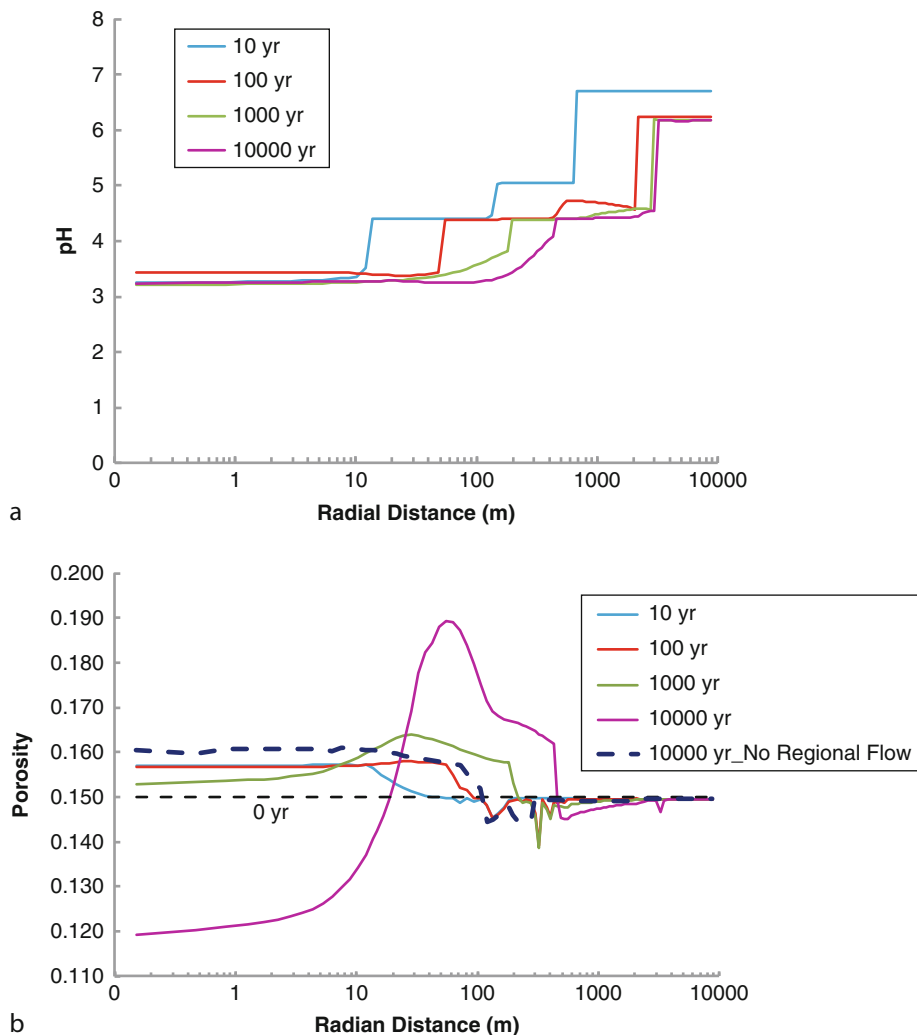
Comparison of flush simulations with geochemical and clone library data from field natural gradient experiment of Anderson et al. [51]. Figure 3 of Istok et al. [43]

the feedback of dissolution – precipitation reactions and changes of permeability and flow patterns – and thermal and mechanical stress. Applications of CRMT include fate and transport of metal and radionuclides in

groundwater systems, geologic carbon sequestration, sediment diagenesis, geothermal energy exploration and production, sea water intrusion, and formation of ore deposits.

As an example, Liu et al. [48] used multiphase reactive flow and transport modeling to simulate large-scale CO₂ injection (a million tons per year for 100 years) into Mt. Simon sandstone, a major candidate saline reservoir in the Midwest of USA. The long-term fate of CO₂ was simulated by extending the modeling period to 10,000 years (the predictive utility of geochemical modeling). The results indicate that most of the injected CO₂

remains within a radius of 3,300 m lateral distribution. Four major trapping mechanisms and their spatial and temporal variations are evaluated in the simulations: hydrodynamic, solubility, residual, and mineral trapping. A strongly acidified zone (pH 3–5) forms in the areas affected by the injected CO₂ (0–3,300 m), and consequently causes extensive mineral precipitation and dissolution (Fig. 3).



Geochemical Modeling in Environmental and Geological Studies. Figure 3

(a) Simulated pH variations as a function of radial distance at year 10, 100, 1,000, and 10,000 with regional flow rate of 0.3 m/year. The initial pore fluids had a pH of 6.9. (b) Simulated porosity variations as a function of radial distance at year 10, 100, 1,000, and 10,000 with a regional flow rate of 0.3 m/year, compared at year 10,000 with no regional flow

Coupled reactive mass transport models are great tools to further understand geochemical reaction networks. Typically, CRMT modeling generates a great amount of numerical experimental data. Significant amounts of time and energy are necessary to dissect and distill the information on what reactions have happened and how reactions are coupled.

Future Directions

There is no doubt that geochemical modeling plays an increasingly important role in environmental sustainability sciences and technology, as environmental decisions (carbon dioxide storage, remediation) are made partly based on model predictions. Computer code development has given flexibility to most applications. However, modelers need to have a grasp of the fundamental understanding of the underlying sciences and the limitations of geochemical models.

Zhu [9] gave a list of research needs to model geochemical reactions:

1. Internally consistent standard state thermodynamic properties for minerals, particularly minerals with complex and variable chemical compositions and structures like smectite. More accurate thermodynamic properties for common minerals like feldspars would help to resolve the controversy on the Al-bearing minerals.
2. More experimental data and resolution of ambiguities surrounding the speciation of aqueous elements like that for Al species.
3. Improvements in sampling and filtration of natural and laboratory samples for better saturation state assessments.
4. Solid solution models for feldspar and clay minerals.
5. More measurements of rate – free energy of reaction relations at a variety of temperature and pH conditions, leading to accurate theoretical or empirical correlations.
6. Rates and rate laws for precipitation reactions, and improved understanding of nucleation process.
7. Pitzer activity coefficient parameters for trace elements and for all elements at elevated temperatures.
8. More rigorous treatment of experimental data with statistical analysis.
9. Assessment of error propagations.

Acknowledgment

The writing of this entry was also made possible with continued financial support from the US National Science Foundation (EAR0423971, EAR0509775, EAR 0809903) and the US Department of Energy (DE-FG26-04NT42125, DE-FE0004381). Any opinions, findings, and conclusions or recommendations expressed in this material, however, are those of the authors and do not necessarily reflect the views of the US Government or any agency thereof.

Bibliography

Primary Literature

1. Zhu C, Anderson GM (2002) Environmental applications of geochemical modeling. Cambridge University Press, London, p 304
2. Helgeson HC et al (1970) Calculation of mass transfer in geochemical processes involving aqueous solutions. *Geochim Cosmochim Acta* 34:569–592
3. Fang YL, Yeh GT, Burgos WD (2003) A general paradigm to model reaction-based biogeochemical processes in batch systems. *Water Resour Res* 39(4):1083
4. Nordstrom DK (2007) Modeling low-temperature geochemical processes. In: Drever JI (ed) *Surface and ground water, weathering and soils, treatise on geochemistry*. Elsevier, New York, pp 1–38, online update
5. Wolery TJ (1992) EQ3/6, A software package for geochemical modeling of aqueous systems: package overview and installation guide (Version 7.0). URCL-MA-110662-PT-I, University of California/Lawrence Livermore Laboratory, California/Livermore, p 41
6. Kharaka YK et al (1988) SOLMINEQ.88: a computer program for geochemical modeling of water-rock interactions. *Water-resources investigations report 88–4227*, US Geological Survey
7. Parkhurst DL, Appello AAJ (1999) User's guide to PHREEQC (Version 2)-a computer program for speciation, batch-reaction, one dimensional transport, and inverse geochemical modeling. *Water-resource investigation report*, US Geological Survey, p 312
8. Allison JD, Brown DS, Novo-Gradac KJ (1991) MINTEQA2/PRODEFA2, a geochemical assessment model for environmental systems, Version 3.0 user's manual
9. Zhu C (2009) Geochemical modeling of reaction paths and geochemical reaction networks. In: Oelkers EH, Schott J (eds) *Thermodynamics and kinetics of water-rock interaction*. Mineralogical Society of America, Washington, pp 533–569
10. Johnson JW, Lundeen SR (1994) GEMBOCHS thermodynamic data files for use with the EQ3/6 software package. Lawrence Livermore National Laboratory, p 99
11. Johnson JW, Oelkers EH, Helgeson HC (1992) SUPCRT92 – A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species,

- and reactions from 1-bar to 5000-bar and 0°C to 1000°C. *Comput & Geosciences* 18(7):899–947
12. Helgeson HC et al (1978) Summary and critique of the thermodynamic properties of rock forming minerals. *Am J Sci* 278A:569–592
 13. Wagman DD et al (1982) The NBS tables of chemical thermodynamic properties – selected values for inorganic and C-1 and C-2 organic-substances in SI units. *J Phys Chem Ref Data* 11(Supplement 2):392
 14. Berman RG (1988) Internally-consistent thermodynamic data for minerals in the system $\text{Na}_2\text{O}-\text{K}_2\text{O}-\text{CaO}-\text{MgO}-\text{FeO}-\text{Fe}_2\text{O}_3-\text{Al}_2\text{O}_3-\text{SiO}_2-\text{TiO}_2-\text{H}_2\text{O}-\text{CO}_2$. *J Petrol* 29(2):445–522
 15. Berman RG (1990) Mixing properties of Ca-Mg-Fe-Mn garnets. *Am Mineral* 75:328–344
 16. Holland TJB, Powell R (1998) An internally consistent thermodynamic data set for phases of petrological interest. *J Metamorph Geol* 16:309–343
 17. Nordstrom DK et al (1990) Revised chemical equilibrium data for major water-mineral reactions and their limitations. In: Melchior DC, Bassett RL (eds) *Chemical modeling of aqueous systems II*. American Chemical Society, Washington, pp 398–413
 18. Robie RA, Hemingway BS (1995) Thermodynamic properties of minerals and related substances at 298.15 K and 1 bar (10^5 pascals) pressure and at higher temperatures. *US Geological Survey Bulletin* 2131, p 456
 19. Grenthe I et al (1992) *The chemical thermodynamics of uranium*. Elsevier, New York
 20. Helgeson HC, Kirkham DH, Flowers GC (1981) Theoretical prediction of the thermodynamic behavior of aqueous electrolytes at high pressures and temperatures. IV. Calculation of activity coefficients, osmotic coefficients, and apparent molal and standard and relative partial molal properties to 600°C and 5 kb. *Am J Sci* 281:1249–1516
 21. Shock EL, Helgeson HC (1988) Calculation of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: correlation algorithms for ionic species and equation of state predictions to 5 kb and 1000°C. *Geochim Cosmochim Acta* 52:2009–2036
 22. Shock EL, Helgeson HC, Sverjensky DA (1989) Calculations of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: standard partial molal properties of inorganic neutral species. *Geochim Cosmochim Acta* 53:2157–2183
 23. Sverjensky DA, Shock EL, Helgeson HC (1997) Prediction of the thermodynamic properties of aqueous metal complexes to 1000°C and 5 kb. *Geochim Cosmochim Acta* 61(7):1359–1412
 24. Shock EL et al (1992) Calculation of thermodynamic and transport properties of aqueous species at high pressures and temperatures. Effective electrostatic radii, dissociation constants and standard partial molal properties to 1000°C and 5 kb. *J Chem Soc London, Faraday Trans* 88: 803–826
 25. Tanger JC, Helgeson HC (1988) Calculations of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: revised equation of state for the standard partial molal properties of ions and electrolytes. *Am J Sci* 288:19–98
 26. Oelkers EH, Helgeson HC (1990) Triple-ion anions and polynuclear complexing in supercritical electrolyte-solutions. *Geochim Cosmochim Acta* 54(3):727–738
 27. Nordstrom DK, Munoz JL (1994) *Geochemical thermodynamics*, 2nd edn. Blackwell, Oxford
 28. Harvie CE, Moller N, Weare JH (1984) The predication of mineral solubilities in natural waters: the Na-K-Mg-Ca-H-Cl-SO₄-OH-HCO₃-CO₃-CO₂-H₂O system to high ionic strength at 25°C. *Geochim Cosmochim Acta* 48(4):723–751
 29. Plummer LN et al (1988) A computer program incorporating Pitzer's equations for calculation of geochemical reactions in brines. *Water resources investigations report* 88–4153, US Geological Survey, p 310
 30. Xu T et al (2004) TOUGHREACT user's guide: a simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media (V1.2). Paper LBNL-55460. Lawrence Berkeley National Laboratory
 31. Wolery T et al (2004) Pitzer database development: description of the Pitzer geochemical thermodynamic database data0.yppf. Appendix I in *In-Drift precipitates/salts model* (P. Mariner) report ANL-EBS-MD-000045 REV 02. Bechtel SAIC Company, Las Vegas
 32. Garrels RM, Thompson ME (1962) A chemical model for sea water at 25 °degC and one atmospheric pressure. *Am J Sci* 260:57–66
 33. Liu FY et al (2010) Antimony speciation and contamination of waters in Xikuangshan Sb mining and smelting area, China. *Environ Geochem Health*. doi:10.1007/s10653-010-9284-z
 34. Lindberg RD, Runnells DD (1984) Groundwater redox reactions - an analysis of equilibrium state applied to Eh measurements and geochemical modeling. *Science* 225(4665):925–927
 35. Stumm W, Morgan JJ (1996) *Aquatic chemistry, chemical equilibria and rates in natural waters*. Wiley, New York, p 1022
 36. Stumm W (1992) *Chemistry of solid-water interfaces: processes at the mineral-water and particle-water interface in natural systems*, 1st edn. Wiley, New York
 37. Dzombak DD, Morel FMM (1990) *Surface complex modeling: hydrous ferric oxide*. Wiley, New York, 393
 38. Appelo CAJ, Postma D (2005) *Geochemistry, groundwater and pollution*. A. A. Balkema, Leiden
 39. Helgeson HC (1968) Evaluation of irreversible reactions in geochemical processes involving minerals and aqueous solutions-1. Thermodynamic relations. *Geochimica et Cosmochimica Acta* 32:853–877
 40. Helgeson HC (1979) Mass transfer among minerals and hydrothermal solutions. In: Barnes HL (ed) *Geochemistry of hydrothermal ore deposits*. John Wiley & Sons, New York, pp 568–610
 41. Anderson GM, Crerar DA (1993) *Thermodynamics in geochemistry: the equilibrium model*. Oxford University Press, New York, 588
 42. Zhu C et al (2010) Coupled alkali feldspar dissolution and secondary mineral precipitation in batch systems: 4.

- Numerical modeling of kinetic reaction paths. *Geochimica Et Cosmochimica Acta* 74(14):3963–3983
43. Istok JD et al (2010) A thermodynamically-based model for predicting microbial growth and community composition coupled to system geochemistry: application to uranium bioreduction. *J Contam Hydrol* 112(1–4):1–14
 44. Liu C et al (2001) Kinetic analysis of the bacterial reduction of goethite. *Environ Sci Technol* 35(12):2482–2490
 45. Roden EE (2008) Microbiological controls on geochemical kinetics 1: fundamentals and case study on microbial Fe(III) oxide reduction. In: Brantley SL, Kubicki J, White AF (eds) *Kinetics of water-rock interaction*. Springer, New York, pp 335–415
 46. Zhu C (2003) A case against Kd-based transport model: natural attenuation at a mill tailings site. *Comput Geosci* 29:351–359
 47. Yeh GT, Tripathi VS (1989) A critical evaluation of recent development of hydrogeochemical transport models of reactive multi-components. *Water Resour Res* 25(1):93–108
 48. Liu FY et al (2010) Coupled reactive transport modeling of CO₂ Sequestration in the Mt. Simon sandstone formation, Midwest U.S.A. *Int J Greenh Gas Con* 5:294–307
 49. Zhu C (2004) Coprecipitation in the barite isostructural family: 1. Binary mixing properties. *Geochimica et Cosmochimica Acta*, 68(16):3327–3337
 50. Zhu C (2004) Coprecipitation in the barite isostructural family: 2. Binary mixing properties. *Geochimica et Cosmochimica Acta*, 68(16):3339–3349
 51. Anderson TT, Vrionis HA, Ortiz-Bernard I, Resch CT, Long PE, Dayvault R, Karp K, Marutzky S, Metzler DR, Peacock A, White DC, Lowe M, Lovley DR (2003) Stimulating the in situ activity *Geobacter* species to remove uranium from the groundwater of a uranium-contaminated aquifer. *Appl Environ Microb* 69:5884–5891

Geoengineering Policy and Governance Issues

SEAN LOW¹, NIGEL MOORE¹, ZHEWEN CHEN¹,
KEITH McMANAMEN², JASON J. BLACKSTOCK³

¹Centre for International Governance Innovation,
Waterloo, ON, Canada

²University of Waterloo, Waterloo, ON, Canada

³Kennedy School of Government, Cambridge,
MA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

History of Weather Modification and Governance
Frameworks

Geoengineering Technologies and Governance
Challenges

Current Governance Landscape

Future Directions

Bibliography

Glossary

Carbon geoengineering A variant of geoengineering also known as Carbon Dioxide Removal (CDR) concepts aimed at capturing carbon dioxide directly from the atmosphere by either enhancing existing natural sinks or by using chemical engineering technologies.

Geoengineering Also known as “climate engineering” refers to the deliberate and technological manipulation of the climate system to forestall the worst effects of global warming.

Governance The management of political issues and physical systems that relies not only on (traditional) government at state level, but upon a wider range of actors at the international (international organizations, minilateral clubs), substate (provincial and municipal government), and nonstate (industry, civil society, knowledge networks) levels with such management often involving coalitions across multiple levels and actor types.

Solar geoengineering A variant of geoengineering also known as Solar Radiation Management (SRM) concepts aimed at enhancing reflecting incoming sunlight back into space, preventing absorption by Earth’s atmosphere, surface or oceans, and thereby reducing global temperatures.

Weather modification An antecedent to climatic geoengineering focused on the manipulation of local atmospheric conditions to induce short-term, bounded changes in weather. Such methods are different from geoengineering methods in that they are smaller in scale and intent of atmospheric modification.

Definition of the Subject

Geoengineering – the deliberate and technological manipulation of the climate system to forestall the worst effects of global warming (*also referred to as*

climate engineering) – has recently emerged as a novel and controversial issue in climate governance. It is sometimes proposed as an insurance policy, should either (a) primary efforts to develop sustainable energy and societal systems prove unable to quickly enough overcome the inertia of current ones, or (b) uncertainty in the climate system lead to unexpectedly large damage to societies and ecosystems [1]. This entry explores the current attempts and future ideas for governing emerging geoengineering research programs and technologies in ways that effectively manage their climatic and societal impacts.

Introduction

The international and national discourses surrounding geoengineering are currently in an early formative stage, bringing together scientific, ethical, legal, and political dimensions, whose details and interlinkages have been at best only preliminarily mapped to date. Given the potential accessibility and proposed utility of various geoengineering techniques, the discussion of geoengineering is expected to expand and evolve significantly as new actors begin to engage with the issue.

As a result, the core challenge of designing governance frameworks for geoengineering technologies, particularly at the international level, stems from uncertainty about the future. On the scientific side, research and deployment is confronted by large constraints in forecasting impacts and in running meaningful field tests. Political contentions include the potentially unequal distribution of technology ownership; risk in field-testing and deployment; the unclear effects of the debate's politicization by a range of states, researchers, private developers of technology, and civil societies; the potential introduction of moral hazard into climate change mitigations efforts; and complex intersections with a number of global governance issues and policy communities. Underlying each of these political contentions are ethical perspectives on how humanity should interface with the natural world.

Hence, the crux of governance is not only to provide guidelines and structure for managing near-term actors and issues, such as scientists and research agendas. It must also set a framework for managing new related issues and interconnections as they evolve. In this context, caution regarding “locking in”

governance mechanisms designed to manage only near-term issues is needed, as narrow frameworks might hinder capacity to address future possibilities and issues. Even independent of lock in, any regulation enacted in the near future could clearly have a formative impact on the future governance of geoengineering. However, with the current lack of an institutional forum for discussion, national legal frameworks, or even coherent rules and framings among the actors now driving the debate, projecting what that impact might be is exceedingly difficult. For these reasons, recent initiatives – such as the Solar Radiation Management Governance Initiative (SRM-GI) and Asilomar Conference on Climate Intervention Technologies [2, 3] – aiming to promote the development of responsible governance for emerging geoengineering research and technologies have focused first on expanding the conversation to include greater diversity in the perspectives involved.

The following sections will provide a general overview of the various components of the geoengineering governance debate. For more detail on any of the issues raised, readers are encouraged to examine the references herein in detail. The section on “[History of Weather Modification and Governance Frameworks](#)” chronicles historic environmental modification activities as limited antecedents with potential influence on how geoengineering governance may evolve. The section on “[Geoengineering Technologies and Governance Challenges](#)” outlines important distinctions between different geoengineering technologies and stages of research, and highlights how these relate to a range of technical and sociopolitical issues that may require governance. The section on “[Current Governance Landscape](#)” then provides an overview of the currently emerging governance and policy landscapes, including identifying key actors, institutions, and issues that have shaped and are currently shaping the governance landscape thus far. Finally, as the current landscape remains far from comprehensively developed, the last section on “[Future Directions](#)” overviews existing proposals for the future governance and regulation of geoengineering research and technologies.

Given the nascent state of national and international discourses on geoengineering, this entry necessarily focuses as much on speculative future proposals for governance as on existing governance or policy.

This balance highlights the uncertainty – and potential for influencing future developments – that exists within the discourse.

History of Weather Modification and Governance Frameworks

While the current concept of geoengineering is relatively modern, it can be situated within a long tradition of human attempts to influence their environmental surroundings. Throughout history, mankind has imagined and fantasized about weather control; such stories have been found in myths, religious texts, traditional and cultural practices, and science fiction. Granted, the nature of weather control aligns only partially with geoengineering. Yet, its historical treatment has implications for looking at geoengineering from the point of view of antecedent governance efforts.

In practice, the attempt to manipulate atmospheric conditions had its roots in the Enlightenment vision of deconstructing natural processes for the betterment of human society. However, it was not until the nineteenth century that scientific research programs materialized in this capacity. The early weather engineers were rainmakers who proposed various physical, chemical, and electrical approaches to manipulate precipitation.

Prominent figures included James Pollard Espy (1785–1860), a US government appointed meteorologist, who actively promoted the idea of commercial rainmaking by cutting and burning forest woods. Edward Powers and Daniel Ruggles were former US generals who theorized that cloud condensation by explosive agent could lead to rainfall [4]. Other prominent figures of this “concussive” stream included Robert St George Dyrenforth (1844–1910), a lawyer from Washington D.C. who received government funding to lead such experiments, J. B. Atwater, who filed a weather patent to disrupt tornadoes, and Laurice Leroy Brown, who filed a patent application for an “automatic transporter and exploder for explosives aiding rain-fall” [4]. Instead of engaging in serious science, rainmakers of the era often sought to defraud desperate farmers. A famous example was Charles Hatfield, who evaporated chemicals to “coax(ed) rain from the sky” [5]. Projects of the age were backed by poor theory, speculative knowledge, and insufficient financial support; and regulatory frameworks were largely nonexistent.

Whereas agriculture interests were the primary driver for nineteenth-century weather controlling schemes, research into weather control during the two World Wars was supported mostly by military patronage as well as some corporate financial support. Aerial fogs and vapors impeded forward visibility of troops and bombing raids, and the escalation of World War II called for immediate scientific attention. In 1940, Britain created the Petroleum Warfare Department (PWD) and charged it with the task of developing a reliable method of clearing fogs called the Fog Investigation and Dispersal Operation (FIDO), the project brought together scientists, engineers, industrialists, and policy makers under one umbrella. With FIDO, the British and Allied air forces were able to assume normal military activities as the Germans were bound to the ground. Despite its success as a wartime invention, FIDO ultimately proved to be too costly to be commercially viable [4]. However, this complex of interests carried over to the Cold War, with even greater strategic interests of weaponizing weather control techniques.

The scientific foundation began in the General Electric (GE) Research Laboratory, where in 1946, Vincent Schaefer and Irving Langmuir discovered a weather-control technique that used dry ice and silver iodide to “supercool” clouds [4]. Langmuir envisioned that the technology would have the potential to deflect hurricanes, generate large-scale precipitation, and clear the sky for aviation services. The confluence of military interests and GE’s concerns over liability threats prompted the transfer of research to a newly established classified cloud-seeding program called Project Cirrus, a collaboration between GE, the US Air Force, the US Army Signal Corps, and the Office Naval Research. Cirrus resulted in more than 250 trials and experiments during 1947–1952, but was canceled due to tort liability lawsuits [4]. Project Stormfury, successor to Cirrus and operational from 1962 to 1983, was another collaboration between the Weather Bureau and the military in exploring cloud-seeding experimentation.

During the Vietnam War, Operation Popeye, a clandestine field trial of cloud seeding was conducted along the Ho Chi Minh Trail to stonewall traffic. With support from the US administration, a much larger scale program known as Operation Motorpool, covering areas over Laos, Cambodia, North and South Vietnam, ran until July 1972 [6]. Similarly, French

forces also attempted artificial rainmaking to impede the movement of Vietnamese troops; not to mention the frequent practice of cloud seeding by Moscow in the 1950s and the 1960s.

The fact that many of the research and trials remained in the domain of military operations shielded them from public criticism and regulatory oversight. Therefore, despite scant court cases on the tort liability of cloud-seeding experiments, no statutory laws were erected to regulate weather modification during the interwar period. However, when a story about the operations in Vietnam broke in the Washington Post in March 1971, it became a public relations disaster that led to pressure from members of the Congress on the Nixon administration to cease the operations. The efforts to ban environmental warfare were led by Senator Claiborne Pell in the Senate and Representative Donald M. Fraser in the House, leading to the passing of numerous bills in 1975 that sought to prevent the weaponization of weather modification [4].

The Soviet Union was quick to respond to the “Water-gate of weather warfare.” First, it invited the Nixon administration to sign the “Joint Statement Concerning Future Discussion on the Dangers of Environmental Warfare” at the Moscow Summit in 1974 [7]. It then presented the United Nations with a proposal to establish an international convention to outlaw weather modification as a weapon of war. This diplomatic initiative caused the Ford administration to commit the USA to the negotiation and eventual signing of the Convention on the Prohibition of Military or Any Other Hostile Use of environmental Modification Techniques (ENMOD), which banned all militarized environmental modifications with “widespread, long-lasting or severe effects as the means of destruction, damage or injury to any other State Party.” ENMOD went into effect in October 1978 with 70 state members. However, the treaty was viewed as deeply flawed: It contained vague texts, established a high threshold for violation, and did not prohibit research and development in the field, which made the treaty almost unusable [4]. Nonetheless, the ENMOD treaty has still been discussed as potentially applicable to certain geoengineering techniques due to the potential for detrimental transboundary impacts [8–11].

Despite limited scientific evidence of success over the past half-century, the turn of the century still witnessed governments in drought-prone countries

resort to cloud-seeding techniques to enhance precipitation. During the 2008 Summer Olympics, Beijing deployed over 1,000 rain dispersal rockets to prevent rain during the opening and closing ceremonies [12]. Over the summer months of 2010, the Abu Dhabi government commissioned an \$11 M project that used ionizers to generate storms [13].

The boundary between what constitutes weather modification (*a single nation attempting to modify their own weather*) versus geoengineering (*sufficiently large modification of the atmosphere to have significant transboundary impacts*) currently remains ill defined – and, for the moment at least, uncontested. However, the potential for both evolving weather modification and geoengineering experiments to encroach on this boundary points to the importance for effective governance of understanding the above history. In the long tradition of weather modification, human civilization has largely bypassed the ethical, social, and legal dimensions of the issues, rendering much of its governance and legislative aspects unresolved to the present day. Yet, the nature and magnitude of weather modification efforts to date are child’s play compared to the potential for some geoengineering technologies to generate impacts over much larger areas and timescales. Thus, the historical incapacity to regulate atmospheric interference causes concerns that geoengineering research will demonstrate a repeat of the same evolutionary pattern.

Geoengineering Technologies and Governance Challenges

Distinguishing Carbon and Solar Geoengineering

Geoengineering is a blanket term that encompasses a suite of very different techniques for intervening in the climate system [14]. In designing effective governance arrangements for different geoengineering technologies and research, it is therefore crucial to take account of the differences between them. Moreover, as these technological constructs will themselves evolve with time, governance frameworks need to be adaptable to addressing their possible evolution as research progresses in climate science and geoengineering. What follows is a basic description of prominent geoengineering methods that have been envisaged to date, including identification of the various technical and social uncertainties surrounding each.

The broadest differentiation of prominent geoengineering technologies is between *carbon* and *solar* geoengineering, which tackle the climate challenge from very different directions. Through removing excess carbon from the atmosphere, carbon geoengineering (*also called carbon dioxide removal or CDR techniques*) aims to treat the *cause* of climate change by removing carbon dioxide already in the atmosphere. Solar geoengineering (*also known as solar radiation management or SRM techniques*), on the other hand, aims to reflect incoming solar radiation before it is absorbed by Earth's atmosphere, surface or oceans. This strategy aims to offset the energy imbalance created in the Earth's climate by greenhouse gasses trapping outgoing infrared radiation, thereby avoiding or reducing the anticipated "global warming" from that imbalance.

The actual technologies being proposed to implement either carbon dioxide removal or solar radiation deflection vary significantly. Carbon geoengineering, for example, can be broken down into two subcategories: engineered and ecosystem. The prominent techniques in these categories differ in ways that would suggest the requirement of very different arrangements for their effective governance. The most studied engineered technique is *direct air capture*, which would utilize machines (the size of large buildings) that draw in ambient air, chemically remove carbon dioxide from it, and store that carbon somewhere for a very long time [15]. On the other hand, the ecosystem carbon geoengineering technique that has garnered the most attention from scientific and environmental governance communities is ocean fertilization, which is premised on the enhancement of a process of CO₂ removal that already exists in nature. With ocean fertilization, nutrients (of which iron is most often discussed) that limit the growth of organisms such as algae and phytoplankton are added to the oceans to stimulate the growth of these species. These organisms require carbon to live and grow which they fix from the atmosphere, and when they die they sink and eventually sequester the carbon in deep ocean sediments [16].

Solar geoengineering tackles the climate challenge from a different direction, focusing instead on attempting to enhance the albedo of the planet. Notable options for achieving this include launching mirrors into space to reflect sunlight before it reaches the atmosphere, brightening clouds by spraying a mist

of cloud condensation nuclei (fine particles) into the lower atmosphere, and injecting reflective aerosols such as sulfates [17] – or possibly synthetic variants made of nanoparticles [18] – into the stratosphere. Space mirrors remain presently unfeasible at scale, while the most prominent cloud brightening technique would use globally distributed oceangoing vessels to remove sea salt from the water below them and spray it into the air above [19]. Stratospheric aerosols are perhaps the most plausible solar geoengineering method thus far proposed, in part because they mimic a natural process that is known to cool the planet significantly – large volcanic eruptions which send sulfate aerosols high into the stratosphere for a period of about a year. Delivery methods for distributing aerosols in the stratosphere include balloons, guns fired from ground level, and airplanes [1].

At this early stage of technical research and development, it is important that governance frameworks be developed with recognition that the technicalities of hypothesized geoengineering methods will evolve significantly. To date, most concepts remain largely theoretical, with the technologies for deployment not yet having been developed or demonstrated. Over the coming decades there will assuredly be changes to these proposals. Some may be scrapped and forgotten, and new ideas will emerge. As a result, the early arrangements for governance would best be made flexible to these future developments, whatever they happen to be.

Nonetheless, governance assessments of geoengineering have emphasized that uncertainty about future technologies should not be an excuse for complete inaction today, as there is a great deal of information that can be teased out of prior geoengineering research – and relevant precedents for other evolving technologies – that can guide near-term governance efforts [20]. For one, it has become apparent that some techniques could be extremely inexpensive and yet have the ability to significantly impact climate on a global scale [21]. Take, for example, stratospheric aerosol injection, which may be able to cool the planet by multiple degrees centigrade over a timescale of a few years at a cost of a few billion dollars per year [21]. The technologies necessary to develop this capability are relatively simplistic by modern standards, and more than likely could be developed by a small group of actors [22, 23]. On top of this, deployment has the potential to be carried out

clandestinely given the relatively small amount of aerosol loading required and deployment of this kind would impact the globe, with potentially severe side effects for some human and natural systems. Clearly, with this type of scenario, there is a near-term need for suitably tailored governance.

Despite the negative connotations of geoengineering in the previous example, the large consequences that will likely be faced if climate change is under addressed by insufficient mitigation efforts suggests that geoengineering research may be a prudent investment. If this is the case, there may be a need for near-term governance frameworks for geoengineering research to foster open and collaborative research projects aimed at uncovering whether or not there are techniques that might be worthy of deployment should dangerous climate change appear unavoidable [22].

Research Stages

Research into geoengineering, some of which is already ongoing in the absence of such a governance framework, will follow a predictable set of stages beginning with basic theory and modeling [1]. Understanding stages and categories of research, along with challenges and impacts they may pose, could be a starting point for establishing governance by pointing out which types of activities deserve governance attention, which do not, and therefore at what point in the future a legitimate system of research governance must be put in place. See Table 1 for further detail on different categories of research and the potential governance issues they raise.

The first two stages of research do not involve the release of substances into the environment – they involve strictly lab-based activities. Thus, in the eyes of many, governance of these stages is not warranted – any restriction on these activities would not directly protect the environment or citizens from harm, and likely would be very difficult to enforce because geoengineering research at these stages appears (and could even be disguised as) very similar to other kinds of research such as climate science, meteorology, volcanology, etc. However, other perspectives hold that even these early stages of research will have important path-dependent impact of the types of technologies that evolve and are eventually used, and thus need to be incorporated into broader governance frameworks – even if only through broader discussions

of the implications of certain research directions. Moreover, issues of technology ownership and transparency of research activities are directly raised at this stage – both of which are at the core of most evolving geoengineering governance proposals.

The two latter stages of research involve an intentional release of substances into the environment of some kind. Small-scale field studies are defined as having a negligible environmental impact and for this reason there is justification of some reticence regarding strict regulation as these activities would not have transboundary effects, nor would their effects be climatically relevant. Examples might include testing the effectiveness of a nozzle for spraying sulfate particles. To many within the field of geoengineering science, a key governance question is where to define the threshold between negligible and nonnegligible, as this ought to be the point where governance restrictions kick in. Unfortunately, as long as the threshold between negligible and nonnegligible remains difficult to define, this stance toward governance may be problematic to implement.

An alternative viewpoint worthy of mention sees research stages as irrelevant. Proponents of the “slippery-slope” notion hold the view that early stages will increase dramatically the likelihood of eventual deployment and thus are nonnegligible and should themselves be subject to regulation [11]. This argument raises an important set of questions: Is geoengineering research unique? Ought it to be treated like chemical weapons and other kinds of research that have obviously dangerous consequences that make their restriction critical even at early theoretical stages? Can and should geoengineering researchers be trusted? Where in all of this do the widely held value of scientific freedom, and the commitment to the scientific search for truth fit? These are difficult questions that can foster conflict between alternative viewpoints rooted in different ethical perspectives.

Issues

Sociopolitical Linkages Geoengineering requires a highly nuanced governance approach to accommodate the various impacts of research, development, and possible deployment. Unfortunately, thinking of geoengineering in and of itself is not necessarily helpful for governance – in fact, this perspective could be

Geoengineering Policy and Governance Issues. Table 1 Geoengineering research stages

Stage	Description	Challenges	Status	Direct environmental impact
1. Theory and modeling	Publications and computational models studying the anticipated climatic impacts of geoengineering techniques	International cooperation, research funding, and development of more comprehensive models	Studies began more than 20 years ago and continue for both carbon and solar technologies	None
2. Technology development	Design and laboratory development of geoengineering deployment technologies	Emergence of governance issues when technologies are patented or classified. Who has access to and control over new technologies?	Many carbon geoengineering technologies are currently under development	None
3. Subscale field testing	Feasibility testing of geoengineering deployment technologies at levels posing “demonstrably negligible” environmental and transboundary risks	Evaluating risks and modeling uncertainties related to the environmental impacts of field testing	Limited recent tests of atmospheric aerosol injection and ocean iron fertilization have taken place	Small, regional, temporary (negligible)
<i>Difficult to define threshold</i>				
4. Climatic impact testing	Testing the global climate impacts of geoengineering deployment, nominally at scales below actual deployment, but with notable transboundary environmental impacts	Environmental and governance challenges of experiments spread unevenly at local, national, and regional levels. Definition of large scale	A few, limited proposals for this kind of testing to begin soon	Potentially global and large (nonnegligible)

damaging to the design of a successful governance regime because geoengineering is tied inexorably to many other issues. Prudence dictates that these linkages should be accounted for in discussions and deliberations of geoengineering governance challenges.

The climate regime is the most obvious and fundamental decision-making structure with which geoengineering governance must be intimately tied. Further potential need for geoengineering can be seen as proportional to the lag in progress toward mitigation, and related to limited progress toward (or capacity for) adaptation. With their successes and failures, as well as the progression of geoengineering research from theory to practice, the potential future role of geoengineering will become clearer. For this reason it

appears important to meaningfully link decision-making structures that dictate responses to climate change – whether they be mitigation, adaptation, or geoengineering based.

Another importantly linked issue is that of the economic and social development of least developed nations. The largest impacts of both climate change and geoengineering are likely to be felt by the poorest communities of the world, leading to an important question of how their needs should be addressed [24]. Another tightly linked issue area is that of international security, as both climate change and geoengineering research or attempts at deployment could have destabilizing geopolitical effects [25]. Such linked issues – also currently including agriculture and food

security – are likely to only grow as geoengineering becomes a more well-defined proposition. Air pollution could also be a future candidate because the injection of sulfates into the stratosphere is one of the prominent techniques proposed. The line between stratospheric and tropospheric sulfate could blur in the future – in fact some people have already begun to examine the dichotomy between costly decreases of sulfur emissions over the oceans from ships which do little to improve human health, yet increase warming through reducing cooling tropospheric sulfate aerosols in ways that may accelerate the danger to societies and ecosystems from climate change [26]. This suggests there may be legitimate reasons to coordinate air pollution and geoengineering governance regimes in some respects as well.

There are additional questions regarding post-deployment governance that while far off, must be agreed upon well in advance of deployment and almost assuredly before any large-scale field-testing is underway. The tough governance questions raised by the prospect of having winners and losers – some nations that benefit more from a geoengineering intervention than others, to the point where some areas could face serious negative side effects while others reap tangible benefits from climate change avoidance – are prime examples of this.

Ethics Those who decide whether or how to utilize geoengineering in response to climate change will wield a tremendous amount of power. The expansive normative and ethical implications of such decisions are the core reasons why geoengineering is such a controversial topic, even in its current formative stage. Though seemingly unwieldy, such far-reaching questions can be teased into a number of more precise ones for understanding the implications for governance of the ethical issues entangled in the prospect of geoengineering our climate.

One example is the question of informed consent. Geoengineering resembles the concept of an experiment being performed on the entire global population [27]. Much like experimental medicine, it would be performed in order to avoid a larger problem; however, there is potential for severe negative side effects that could be very damaging. In medical scenarios, informed consent from the patient is required to

carry out treatment. However, in the example of geoengineering, this is impossible: One cannot procure informed consent from everybody on the planet. Implementing geoengineering and large-scale tests of geoengineering methods necessitates a decision-making structure that is remarkably inclusive. Unfortunately, the degree of inclusivity required to make such an experiment ethically acceptable will always be a highly controversial question in and of itself.

There is also the question of liability if something goes wrong. In a scenario where unintended negative side effects are felt by one group of stakeholders, there will certainly be finger-pointing. A successful governance regime will have well-defined courses of action to deal with liability and compensation issues in place well in advance of any large-scale testing [11]. Actors currently engaged in geoengineering discussions today should pay close attention to their personal responsibilities and actions, and how they may be relevant to future liabilities. Scientists in the early stages of research who are pushing for field-testing, as well as those who speak out proposing a ban on all research, should acknowledge the influence that such stances can already have in dictating the future of geoengineering research and development.

These examples are a few of the many ethical issues that arise when geoengineering (both research and potential deployment) is considered. They tend to stem from a small group of high-level concerns. One is the issue of intent. Though humans have been changing our environment for centuries, the scale and intent of this incarnation is different, [28] and intentionality in most moral (and legal) constructs brings with it added responsibility and liability for wrongdoing or negligence. Another issue for many who are skeptical of geoengineering's acceptability is that it is a technologically based fix to a problem caused in large part by the use of other powerful technologies. People with this viewpoint see climate change as a problem requiring behavioral changes – not technical solutions – to ultimately solve. Finally, there is the high-level question of who decides [1]. This brings up issues of informed consent, global participation in the decision-making process, the importance of collaboration, the primacy of the interests of the most vulnerable among us, and many others.

Current Governance Landscape

As the subject of climate intervention technology has gained attention within the scientific community, public debates on the topic have grown in their legitimacy and significance. As a result, the constellation of actors involved has also grown to encompass an array of groups and individuals within the public and private sector, government and NGOs, foundations, private philanthropists, and firms [29]. In spite of the diversity of interests at play, most actors to date have presented arguments somewhere in between the extreme poles of either aiming to have geoengineering implemented in some fashion, or aiming to establish a full moratorium on deployment of the technology.

The recent expansion of attention to geoengineering was nucleated by a landmark 2006 article by Nobel Laureate Paul Crutzen [17]. Within a few years, the emerging attention prompted the UK Royal Society and the US National Research Council to explore geoengineering, and issue reports calling for geoengineering research, and a joint statement by the G8 + 5 nations included a call for an international meeting on geoengineering [30–32]. Policy statements calling for research into climate engineering have also emerged from the American Meteorological Society, the American Geophysical Union, and the UK Institution of Mechanical Engineers [33–35]. Each of these documents emphasized the primacy of reducing GHG emissions, but recommended developing conventions for the scientific community and the blueprints for a formal geoengineering governance framework. The Royal Society concluded:

- ▶ Little research has yet been done on most of the geoengineering methods considered, and there have been no major directed programmes of research on the subject. The principal research and development requirements in the short term are for much improved modelling studies and small/medium scale experiments (e.g. laboratory experiments and field trials). Investment in the development of improved Earth system and climate models is needed to enable better assessment of the impacts of geoengineering methods on climate and weather patterns. . . as well as broader impacts on environmental processes [30].

Prompted by the increasing scientific attention to these issues, committees of the UK Parliament and US

Congress released their own reports on geoengineering [36]. Both bodies explored related issues in a series of science and technology hearings on the subject, and reached conclusions endorsing increased research into emerging geoengineering technologies. As an important indicator of concern for the international dimensions of geoengineering research and technologies, the UK Parliament and US Congress organized their hearings jointly, allowing testimony, documents, and reports from each other's hearings to be presented in their own. This is the first, and thus far only, time in history that such an arrangement between the UK and US legislative bodies has been used.

Government funding for geoengineering research is currently scarce in the US, where scientists rely primarily on private philanthropy and redirected federal research grants. Some researchers have argued that a delay in establishing a federal program will make it progressively harder for the US government to guide these efforts in the public interest as the dialogue continues to move rapidly forward [37]. Elsewhere, public funding has been gradually forthcoming. Both the UK and the European Union have recently provided preliminary grants to a few moderate research projects [38–40]. Small, government-funded projects using computer models are also underway in Germany, [41] and Russia has already conducted at least one geoengineering field test [41]. Nevertheless, a document from the UK Parliamentary Office of Science and Technology conceded:

- ▶ There is currently very little public funding specifically earmarked for geo-engineering. Despite a US Department of Energy White Paper (Unpublished) that in 2001 recommended a \$64 M, five-year programme, less than \$1M of public money is currently directly funding geoengineering research in the USA. In the UK, the Engineering and Physical Sciences Research Council (EPSRC) has proposed a £3M “Ideas Factory” commencing in 2010. To date, therefore, most research has been either funded using existing climate science grants or has been unfunded, performed in researchers' spare time [42].

The same report posited that an international research program of \$100M would significantly increase the scientific and engineering knowledge, as well as provide greater understanding of the risks

associated with altering climate system. (Also in March 2009, the UK's Royal Society proposed that a \$200 M international fund be established for research into geoengineering [43].)

Much of the present funding for geoengineering research comes from the private sector. Leading organizations such as Environmental Defense Fund, Novim Group, and Climate Response Fund rely on the philanthropy of private donors. Even billionaires Bill Gates and Richard Branson have also brought money to bear on geoengineering research. (Since 2007, Gates has put at least \$4.5 M into the Fund for Innovative Climate and Energy Research [44, 45] and Branson offered a \$25 M cash prize rather than a research grant, which went unendowed [46]). To date, despite significant concerns that vested fossil fuel interests may back geoengineering technology development, there remains no evidence that corporations and industries standing to benefit from continued GHG emissions are investing publically in geoengineering research [47, 48].

A group of corporate actors did emerge in the mid-00s with the goal of generating carbon credits to sell from using ocean-fertilization techniques. Ocean fertilization company Planktos' plan to sequester carbon dioxide through a release of iron filings into the Pacific Ocean was blocked by a petition to the EPA by environmental groups invoking the Marine Protection, Research, and Sanctuaries Act (the "Ocean Dumping Act"). Though Planktos attempted to flout the regulation by using a vessel flying under a different national flag, a lack of investors forced the company to abandon the project [49].

As a consequence of such activities, the International Maritime Organization (IMO) became engaged in the governance of iron fertilization projects. (In July 2007, EPA dispatched a memo to the IMO revealing that Planktos intended to proceed with its planned project without the permit required under the Marine Protection, Research, and Sanctuaries Act (the "Ocean Dumping Act"). The document informed IMO that Planktos, opting not to fly a U.S. flag, would be able to avoid U.S. regulations. EPA with no power to regulate advised the member states of the London Convention to carefully evaluate Planktos' plans.) Following an IMO statement issued in 2007 and a report on iron fertilization drafted in 2008, a resolution was made in 2008 at

the London Protocol and Convention on the Prevention of Marine Pollution by Dumping of Wastes and Other Matter that "ocean fertilization activities, other than legitimate scientific research... should be considered as contrary to the aims of the Convention and Protocol and do not currently qualify for any exemption from the definition of dumping" [50–52].

To assess the research and governance challenges of geoengineering, the Climate Response Fund (CRF) organized a conference of nearly 200 experts in various scientific and policy disciplines, which gathered at historically resonant Asilomar in 2010 (the site of a prior 1970s conference on genetic recombination). Created in 2009, CRF was established in order to fund geoengineering research projects and work with national and international partners to communicate information about geoengineering research (*referred to as climate intervention research*) to interested groups and the general public. Asilomar conferees set out to develop a set of voluntary guidelines, or best practices, for the least harmful and lowest risk conducting of research and field-testing [53]. The meeting brought together social and natural scientists to deliberate about geoengineering governance, in a tentative first step toward international dialogue between scientists and nonscientists on principles for future research [54, 55].

The Royal Society also maintains a strong leadership role in the continuing climate engineering discourse. Building from its 2009 report, and in partnership with the Academy of Sciences for the Developing World and Environmental Defense Fund, the Royal Society launched the SRM Governance Initiative (SRM-GI) in March 2010. It aims to develop regulatory frameworks and best practices for the research and possible deployment of SRM technology [2]. SRM-GI remains as an advisory body for international organizations and national governments engaging in the debate, bringing together scientific and policy experts to provide guidance for the conduct of research, and incorporating input from a number of civil society stakeholders.

Opponents to geoengineering have also put forth initiatives to try and prohibit or otherwise hinder its advance. The Action Group on Erosion, Technology and Concentration (ETC Group), a Canadian NGO,

has led the organization of the “Hands Off Mother Earth campaign” for a moratorium on real-world geoengineering experiments or deployment (The ETC group turned down its invitation to the Asilomar Conference on Climate Intervention Technologies and attempted to mobilize opposition against the summit. ETC Group, “Open Letter to the Climate Response Fund and the Scientific Organizing Committee,” <http://www.etcgroup.org/en/node/5080>) [56]. This campaign lobbied the Convention on Biodiversity at its 10th Conference of Parties in 2010, contributing to a decision establishing qualifications on the scope of field tests and deployment [57]. The nonbinding resolution stated that:

- No climate-related geo-engineering activities that may affect biodiversity take place, until there is an adequate scientific basis on which to justify such activities and appropriate consideration of the associated risks for the environment and biodiversity and associated social, economic and cultural impacts, with the exception of small scale scientific research studies that would be conducted in a controlled setting in accordance with Article 3 of the Convention, and only if they are justified by the need to gather specific scientific data and are subject to a thorough prior assessment of the potential impacts on the environment [58].

Meanwhile, the Intergovernmental Panel on Climate Change in 2011 held scoping meetings for the inclusion of geoengineering in its 2013–2014 Fifth Assessment Report (AR5).

Future Directions

The attempts of a nascent landscape of governance actors to assess and regulate geoengineering (see the section on “[Geoengineering Technologies and Governance Challenges](#)”) faces a complex of ethical, social, and political concerns and the diversity of potential future technological options (see the “[Introduction](#)” section). This section outlines a variety of proposed and potential governance options, as well as their capacity and adaptability to confront current and future issues. Some of these proposals – such as the SRM Governance Initiative (SRM-GI) fostering broader international dialogue and decisions at the CBD and London

Convention, both discussed above – are already in play, although their longevity and impacts on governance remain uncertain. Most options remain proposals as yet unexplored in depth. Although there are many ways to categorize the range of proposals, most assessments have tended to characterize governance mechanisms primarily by the level and locale of authority [9–11, 59]. Accordingly, discussion of proposed governance frameworks here are divided into the categories of: (1) Nonstate, (2) National and Minilateral, and (3) International, or treaty-based varieties.

Nonstate Frameworks

These tend to emphasize nonbinding guidelines and other forms of soft law, and possess a degree of autonomy from government regulation. They are also intended to engage a transnational range of nonstate actors: researchers and private sector developers of technology, with varying degrees of public participation and feedback. Proposals range in regulatory strength and organizational complexity from an absence of regulation, to informal and ad hoc peer regulation, to professional or institutional codes of conduct. For the lattermost option, it has been suggested that research standards could be cohered by a network of research institutes or an international geoengineering research organization [59].

Variations of self-regulating standards, albeit with broad societal input, have traction among many early proponents of further research, and have also received support from more reserved actors that nonetheless see the need for further exploration of the technology and its attendant issues and impacts. A number of the leading scientific voices driving the development of geoengineering technologies and debate were key participants in the Royal Society’s SRM-GI and Asilomar II (David Keith and John Shepherd are two examples pointed out by Banerjee [9] p20). The Royal Society’s 2009 report recommends that it should itself partner with international scientific bodies to derive best practices and transparency mechanisms for research [30]. Meanwhile, Asilomar II was itself modeled on the first Asilomar meeting’s normative thrust toward self-regulation [53].

Among social scientists, Victor argues that bottom-up norms and assessments formed by the collaborations of international science organizations and research groups actively engaged in the debate represent a more flexible and evolutionary governance mechanism than existing international treaties. Such an option might be “an active geoengineering research programme, possibly including trial deployments, that is highly transparent and engages a wide range of countries that might have (or seek) geoengineering capabilities.” Victor adds that “[s]imilar approaches have been followed in other international scientific collaborations that have had potentially hazardous side effects, such as the European Organization for Nuclear Research (CERN) and the Human Genome Project” [59].

The benefits of such codes of conduct center around their relevance to the immediate activities of the geoengineering research community. Most notably, such codes of conduct enable responsible exploration of scientific questions that require significant research in order to guide the development of responsible governance frameworks. Moreover, as they would be driven by a more homogeneous range of actors and could potentially remain free of geopolitical motives, self-derived and imposed codes of conduct for research would likely be much easier to negotiate and swifter to implement through transnational scientific institutions. Finally, soft guidelines may have great flexibility to adapt to evolutions in the debate than rigorously negotiated international agreements [59].

There are, however, drawbacks to this approach. Depending on the extent and severity of the code, a voluntary system raises the important issue of noncompliance, particularly by actors operating outside of publically peer-reviewed scientific literature – such as private actors or clandestine government programs. Moreover, while many researchers support some form of public input, it is uncertain what effect the inclusion of more oppositional civil society groups might have on a code of principles, or on the subsequent desire of researchers and technology developers to be restricted by them. Nonetheless, if research codes forego public participation or government oversight, there may be a backlash that questions the legitimacy of the process. Codes

of conduct are also most amenable to the current phase of laboratory research, and potentially – though ambiguously – in small-scale field-testing, when impacts and controversies are limited and more easily defined. However, they will almost certainly prove less effective or relevant to large-scale testing or deployment, as the entry of governmental agendas into governance may render self-regulation obsolete. Finally, there is the basic question of whether some of the actors most in favor of geoengineering should be allowed to formalize the first principles on its regulation.

National and Minilateral Frameworks

The potential for national regulatory frameworks has received the least attention in both academia and policy prescription. Hypothetically, the accepted legal mandate and enforcement capacities (within its own borders) of developed states makes governmental regulation a strong candidate for creating legally binding strictures at all stages of technology development. Besides national-level legislation or regulation, governmental legal frameworks could take a variety of forms. Within federal structures, subnational levels of government (municipal or provincial, with networks thereof) could enact legislation at a level below the state, and might feasibly network across national borders. There is also the possibility for states to cooperate on geoengineering development and regulation on a minilateral basis – a club-based approach that eschews the complications of universal participation in international frameworks.

Few of these options have been discussed in detail, likely because governments have thus far been hesitant to take strong positions on an issue whose controversies – and resulting political blowback – cannot be accurately predicted [47]. SRM-GI and Asilomar II have both noted the need for governments to scope their positions on geoengineering, and the eventual need for state-led oversight mechanisms [60]. Although no governments have formulated a clear position, the initial scoping efforts by the USA and the UK demonstrate the possibility that domestic legal frameworks will be developed. (See, for example, [61]. For example of hearing, see [62] or [63]). Besides

the creation of novel legislation, Hester notes that existing national environmental and air pollution laws and agencies, such as the Clean Air Act and the Environmental Protection Agency in the USA, might be reoriented to regulate the emissions of materials upon which geoengineering initiatives depend (e.g., sulfate aerosols) [49].

A government-focused array of prescriptions could capitalize on the ability of developed states with strong bodies of environmental law and regulatory capacities to create and enforce geoengineering legislation. Such an approach would also uphold the principle of sovereignty, potentially avoiding the problems inherent in determining the extent and form of collective action. Initial actions taken by states also could serve as testing grounds for regulatory measures, societal debate, and building blocks for the negotiation of a more comprehensive framework at the international level. Even if the state proves hesitant to regulate geoengineering, action could also begin at the subnational levels in federal structures.

However, the efficacy of state regulation also relies on the strength and coherence of its own structures. While states of the OECD could legitimately enact strong rulings, it is difficult to forecast how many capacity-deficient states would be able to follow suit. Federalism may be a double-edged sword, as conflicts of interest between the national and subnational levels of government may emerge. Moreover, there are many issues buried in the logic of unilateralism, especially in the development of novel and risky technologies. If states move actively to develop and deploy geoengineering technologies without multilateral consultation, a geoengineering “race” could begin. This could in turn entrench technological know-how, with first-movers maintaining a lead on latecomers. It is possible that an incentive to act early on technology development and to maintain a technological advantage would emerge, and that contestations over intellectual property rights and ownership would follow. Finally, the uncertainty of the transboundary impacts of large-scale field-testing or deployment makes governance options that rely on the good intent of sovereign states, at the least, a controversial prospect [22, 64, 65]. (It is worth, however, noting the arguments of Horton [69], that the dangers of unilateralism are

overstated, and that there are rational incentives for states to collaborate on geoengineering development and regulation.)

International Frameworks

Existing analyses of and proposals for international governance of geoengineering have tended to focus on ways existing international environmental agreements (IEAs) and organizations (IOs) might be “co-opted” to address the core governance issues. These explorations have focused particularly on IEAs and IOs whose mandates either regulate materials that are components to forms of geoengineering, or impact geoengineering more broadly.

A number of academic studies have mapped out the potential intersections of geoengineering regulation with the mandates and the governing capacities of numerous IEAs [9, 11, 20, 47, 59, 67]. There is a general consensus that while no single IEA has a direct mandate for geoengineering or could govern it in light of all its stages of development and interdependent issues. On the other hand, many studies name treaties or organizations that could assert a regulatory jurisdiction over some aspect of the technology, or some stage of the research to deployment process. For example, stipulations of the Convention on Long-Range Transboundary Air Pollution and the International Marine Organization on sulfur emissions limits might be used to regulate solar engineering testing or deployment [10, 11]. The Montreal Protocol might be called upon to fulfill a similar function should sulfate be proven to be an ozone-depleting substance. (Tilmes et al. [71] demonstrated with laboratory research that sulphates deplete the ozone layer, but corroborations are ongoing and the item has not made it onto the agenda of the Montreal Protocol.) For carbon engineering, the Law of the Sea convention might similarly be used to regulate ocean fertilization, as the London Convention and Protocol recently did [10, 11].

IOs and regimes that might address geoengineering more generally range from the UNFCCC (*as the default locus of any debate about geoengineering as a supplement to the climate issue*), to the Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques (ENMOD), a treaty of the Cold War era that prohibits environmental

modification for military or hostile purposes. (Lin [66] notes both, with a primary focus on the UNFCCC. Banerjee [8, 9] is in favor of rebooting ENMOD's mandate to ban geoengineering as a reconceptualized form of environmental modification. MacCracken [70], on the other hand, advocates modifying ENMOD to permit certain forms and amounts of geoengineering. A host of other IOs have been cited, some quite obscure: the Outer Space Treaty, the Antarctic Treaty System, etc. See [10, 11, 20].) The Convention on Biological Diversity's 2010 decision on geoengineering field-testing was a seminal event. Although of uncertain legitimacy, efficacy, or longevity, the CBD's actions may set a precedent for other IOs to make decisions on geoengineering, with varying intents and governing capacities.

Although little discussed, a new regime might be able to govern geoengineering as a holistic issue. As a blank slate, a novel mechanism could be specifically tailored to the issues and stages unique to geoengineering. However, the international arena already suffers from a glut of IOs and IEAs [68], and the creation of a new body might create jurisdictional overlaps and competition for visibility with existing ones. Moreover, the creation and implementation of a new regime with broad participation and legitimacy is a gradual process in and of itself. Combined with existing disagreements over the myriad and evolving facets of geoengineering, as well as the near absence of governmental positionings, a drawn-out time lag for creation may make it difficult for a new multilateral body to be a source of strong regulation in the near future.

The benefits of using co-opted regimes stem from leveraging frameworks with strong capacities and legitimacy, where governance – even if it targets only one facet of the geoengineering issue – might be more quickly enacted. On the other hand, a patchwork of co-opted regimes may reveal regulatory gaps that cannot be filled, or overlaps and conflicts that cannot be mediated. A multiplicity of governance forum with no key institutional home could also create an incentive for proponents or opponents of geoengineering to forum shop at the body most amenable to their interests, creating the potential for conflicting bodies of international law. Finally, existing institutions contain established mandates, organizational structures, and

political logics that inevitably calcify over time. As such, they may have neither the desire nor flexibility to adjust to an issue as novel and complex as geoengineering [20].

Bibliography

1. Blackstock JJ, Battisti DS, Caldeira K et al (2009) Climate engineering responses to climate emergencies, Novim initial study on geoengineering (Novim Study Group 01, 2009). http://www.novim.org/index.php?option=com_contentview=articleid=31:climate-engineering-responses-to-climate-emergenciescatid=1:recent-papersItemid=2
2. SRM-GI details its mission statements on its website at: <http://www.srmgi.org/project-overview/>
3. Asilomar Scientific Organizing Committee (ASOC) (2010) The asilomar conference recommendations on principles for research into climate engineering techniques. Climate Institute, Washington, DC, 2006
4. Fleming JR (2010) Fixing the sky: the checkered history of weather and climate control. Columbia University Press, New York
5. Goodell J (2010) How to cool the planet: geoengineering and the audacious quest to fix earth's climate. Houghton Mifflin Harcourt, Boston
6. Fleming JR (2006) The pathological history of weather and climate modification: three cycles of promise and hype. *Hist Stud Phys Biol Sci* 37(1):3–25
7. Fleming JR (2007) The climate engineers. *Wilson Quart* 31 (Spring):46–60
8. Banerjee B (2010) ENMOD squad: could an obscure treaty protect developing countries from geoengineering gone wrong? *Slate*, 23 Sep 2010
9. Banerjee B (2011) The limitations of geoengineering governance in a world of uncertainty. *Stanf J Law Sci Policy* IV:15–36
10. Bodansky D (1996) May we engineer the climate? *Clim Chang* 33(3):309–321
11. Virgoe J (2009) International governance of a possible geoengineering intervention to combat climate change. *Clim Chang* 95(1–2):103–119
12. Xinhuanet, Beijing disperses rain to dry Olympic night. http://news.xinhuanet.com/english/2008-08/09/content_9079637.htm. Accessed 10 May 2011
13. Leigh K, Abu Dhabi-backed scientists create fake rainstorms in \$11m project. <http://www.arabianbusiness.com/abu-dhabi-backed-scientists-create-fake-rainstorms-in-11m-project-371038.html>. Accessed 10 May 2011
14. Lenton TM, Vaughan NE (2009) The radiative forcing potential of different climate geoengineering options. *Atmos Chem Phys Discuss* 9(1):2559–2608
15. Keith DW, Heidel K, Cherry R (2009) Capturing CO₂ from the atmosphere: rationale and process design considerations. In: Launder B, Thompson MT (eds) *Geo-engineering climate change: environmental necessity or Pandora's box?*

- Cambridge University Press, Cambridge/New York, pp 107–126
16. Boyd P (2004) Ironing out algal issues in the southern ocean. *Science* 304(5669):396–397
 17. Crutzen PJ (2006) Albedo enhancement by stratospheric sulfur injections: a contribution to resolve a policy dilemma? *Clim Chang* 77(3–4):211–219
 18. Keith DW (2010) Photophoretic levitation of engineered aerosols for geoengineering. *Proc Natl Acad Sci USA* 107(38):16428–16431
 19. Salter S, Sortino G, Latham J (2008) Sea-going hardware for the cloud albedo method of reversing global warming. *Philos Trans Roy Soc A Math Phys Eng Sci* 366(1882):3989–4006
 20. Blackstock J, Ghosh A (2011) “Does geoengineering need a global response- and of what kind?” Working paper of the solar radiation management governance initiative meeting, Kavli, pp 1–35, 21–24 Mar 2011
 21. McClellan J, Sisco J et al (2010) Geoengineering cost analysis, contracted engineering cost analysis. Aurora Flight Services, Cambridge. <http://people.ucalgary.ca/~keith/Misc/AuroraGeoReport.pdf>. Accessed 30 Oct 2010
 22. Blackstock JJ, Long JCS (2010) The politics of geoengineering. *Science* 327(5965):527
 23. Victor DG, Morgan GM, Apt J, Steinbruner J, Ricke K (2009) The geoengineering option. *Foreign Aff* 88(2):69–76
 24. Suarez P, Blackstock J, Van Aalst M (2010) Towards a people-centered framework for geoengineering governance: a humanitarian perspective. *Geoeng Quart* 1(1):2–4
 25. Cascio J (2008) Battlefield earth. *Foreign Policy*. http://www.foreignpolicy.com/articles/2008/01/27/battlefield_earth. Accessed 1 July 2011
 26. Morton O (2009) The international maritime organisation’s plans to warm the world. *Heliophase*. <http://heliophase.wordpress.com/2009/08/20/the-international-maritime-organisations-plans-to-warm-the-world/>. Accessed 20 Aug 2009
 27. Morrow DR, Kopp RE, Oppenheimer M (2009) Toward ethical norms and institutions for climate engineering research. *Environ Res Lett* 4(4):045106
 28. Keith DW (2000) Geoengineering the climate: history and prospect. *Annu Rev Energy Environ* 25(1):245–284
 29. “Lift Off.” *Economist* 4 Nov 2010
 30. Shepherd J et al (2009) Geoengineering the climate: science, governance and uncertainty. The Royal Society, London
 31. America’s Climate Choices: Panel on Advancing the Science of Climate Change (2010) Advancing the science of climate change. The National Academies Press, Washington, DC
 32. The National Academies (2008) Joint statement on climate change from G8 + 5 national science academies: climate change adaptation and the transition to a low carbon society. <http://www.nationalacademies.org/includes/G8+5energy-climate08.pdf>. Accessed 1 July 2011
 33. American Meteorological Society Council (2009) AMS policy statement on geoengineering the climate system. http://www.ametsoc.org/policy/2009geoengineeringclimate_amss_tatement.pdf. Accessed 1 July 2011
 34. American Geophysical Union (2009) Position statement: geoengineering the climate system. http://www.agu.org/sci_pol/positions/geoengineering.shtml. Accessed 1 July 2011
 35. Institution of Mechanical Engineers (2009) Geoengineering: giving us the time to act. http://www.imeche.org/Libraries/Key_Themes/IMechEGeoengineeringReport.sflb.ashx. Accessed 1 July 2011
 36. Science and Technology Committee (2010) The regulation of geoengineering (House of Commons, 2009–2010); Rep. Bart Gordon, Engineering the climate: research needs and strategies for international coordination (House of Representatives, 2010)
 37. Caldeira K, Keith DW (2010) The need for climate engineering research. *Issues Sci Technol* 26(1):57–62
 38. Implications and risks of engineering solar radiation to limit climate change. <http://implicc.zmaw.de/>. Accessed 1 July 2011
 39. Integrated assessment of geoengineering proposals. <http://iagp.ac.uk/about-iagp>. Accessed 1 July 2011
 40. Engineering and Physical Sciences Research Council, “Details of Grant Ep/I-1473x/1.” <http://gow.epsrc.ac.uk/Viewgrant.aspx?GrantRef=EP/I01473X/1>. Accessed 1 July 2011
 41. Izrael YA et al (2009) Field experiment on studying solar radiation passing through aerosol layers. *Rus Meteorol Hydrol* 34(5):265–273
 42. Parliamentary Office of Science and Technology (2009) Geoengineering research. <http://www.parliament.uk/documents/post/postpn327.pdf,1>
 43. Royal Society (2009) Geoengineering the climate : science, governance and uncertainty. The UK Royal Society, London. royalsociety.org/Geoengineering-the-climate/
 44. Fund for Innovative Climate and Energy Research, <http://people.ucalgary.ca/~keith/FICER.html>
 45. Kintisch E (2010) Bill Gates funding geoengineering research. *ScienceInsider*, 26 Jan 2010. <http://news.sciencemag.org/scienceinsider/2010/01/bill-gates-fund.html>
 46. Kanter J (2010) Cash prize for environmental help goes unawarded. *New York Times*, 21 Nov 2010. <http://www.nytimes.com/2010/11/22/business/energy-environment/22green.html>
 47. Reynolds J (2011) The regulation of climate engineering. *Law Inn Technol* 3(1): 113–136 http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1813965
 48. Daly H (2011) Geo-engineering or cosmic protectionism? *Daly news*, Centre for the Advancement of the Steady-State Economy <http://steadystate.org/geo-engineering-or-cosmic-protectionism>. Accessed 31 Aug 2011
 49. Hester T (2011) Remaking the world to save it: applying U.S. environmental laws to climate engineering projects, SSRN eLibrary. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1755203. Accessed 1 July 2011
 50. Intergovernmental Oceanographic Commission (2008) Report on the IMO London Convention Scientific Group meeting on ocean fertilization. Intergovernmental Oceanographic Commission (of UNESCO), Paris. http://www.jodc.go.jp/info/ioc_doc/INF/160478e.pdf

51. Resolution LC-LP.1 (2008) see: <http://climate-l.iisd.org/news/imo-london-convention-parties-agree-on-moratorium-on-ocean-fertilization/>
52. IMO (2010) Assessment framework for scientific research involving ocean fertilization agreed. Press release. <http://www.imo.org/mediacentre/pressbriefings/pages/assessment-framework-for-scientific-research-involving-ocean-fertilization-agreed.aspx>.
53. Leinen M (2011) The asilomar international conference on climate intervention technologies: background and overview. *Stanf J Law Sci Policy* IV:1–5. http://www.stanford.edu/group/sjls/cgi-bin/users_images/pdfs/61_Leinen%20Intro%20Perspective%20Final.pdf. Accessed 1 July 2011
54. Kintisch E (2010) 'Asilomar 2' takes small steps toward rules for geoengineering. *Science* 328: 22–23. <http://www.sciencemag.org/content/328/5974/22.full>. Accessed 1 July 2011
55. Kintisch E (2010) We all want to change the world. *Economist* 3: 81–82. <http://www.economist.com/node/15814427>. Accessed 1 July 2011
56. Hands Off Mother Earth, "Organizations". <http://www.handsoffmotherearth.org/organisations/>. Accessed 1 July 2011
57. ETC Group (2007) ETC, Gambling with GAIA, ETC Communique. http://www.etcgroup.org/upload/publication/pdf_file/ETC_COP10GeoBriefing081010.pdf. Accessed 1 July 2011
58. Convention on Biological Diversity (CBD) (2010) Biodiversity and climate change draft decision submitted by the Chair of Working Group I, conference of the parties to the convention on biological diversity tenth meeting, Nagoya, 18–29 Oct 2010, Agenda item 5.6. www.cbd.int/doc/meet-ings/cop/cop-10/in-session/cop-10-I-08-en.doc
59. Victor DG (2008) On the regulation of geoengineering. *Oxford Rev Econ Policy* 24(2):325
60. SRMGI report, Asilomar II, Leinert 2011
61. Borenstein S (2009) Obama looks at climate engineering. Associated Press, 8 Apr 2009. www.fas.org/sgp/crs/misc/R41371.pdf
62. Caldeira K (2009) Geoengineering assessing the implications of large scale climate intervention (statement to US House)
63. Lane L (2009) Researching solar radiation management as a climate policy option (statement to US House)
64. Victor DG et al (2009) The geoengineering option: a last resort against global warming? *Foreign Aff* 88:64–76
65. Robock A (2008) 20 reasons why geoengineering may be a bad idea. *Bull At Sci* 64:14–17
66. Lin AC (2009) Geoengineering Governance. *Issues Leg Scholarsh* 8 (3) <http://www.bepress.com/ils/vol8/iss3/art2>
67. Rayner S (2011) Climate geoengineering governance. *Jahrbuch Ökologie in Press*, Stuttgart, Germany. http://www.jahrbuch-oekologie.de/aktuelles_rayner.htm
68. Keohane RO, Victor DG (2010) "The regime complex for climate change," Discussion paper 2010–33, Harvard Project on International Climate Agreements, Cambridge, MA, p 5
69. Horton JB (2011) Geoengineering and the myth of unilateralism: pressures and prospects for international cooperation. *Stanf J Law Sci Policy* IV
70. MacCracken MC (2006) Geoengineering: worthy of cautious evaluation? *Clim Change* 77(3–4)
71. Tilmes S, Rolf M, Ross S (2008) The sensitivity of polar ozone depletion to proposed geoengineering schemes. *Science* 320:5880

Geologic Carbon Sequestration: Sustainability and Environmental Risk

CURTIS M. OLDENBURG

Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Article Outline

Glossary

Definition of Subject and Its Importance

Introduction

Carbon Dioxide (CO₂) Capture and Storage (CCS):

How Does it Work?

Opportunity and Capacity

Potential Impacts

Potential Impacts to Potable Groundwater

Induced Seismicity

Future Directions

Acknowledgments

Bibliography

Glossary

Carbon dioxide capture and storage (CCS) The capture of CO₂ from fossil-fuel power plants and other industrial point sources and its injection through wells into deep geologic formations for permanent storage.

Consequence An impact arising from the occurrence of an event or process. For example, the consequence of high CO₂ concentrations in the atmosphere is global warming.

Geologic carbon sequestration (GCS)=geologic CO₂ storage (GCS) The last step of CCS in which CO₂ is injected through wells into deep subsurface formations for permanent storage.

Hazard A potential impact or consequence of an event or process. For example, high CO₂ emissions are a hazard to climate because CO₂ is a greenhouse gas.

Likelihood The probability or degree of potential for an event or process to occur. For example, the likelihood of CO₂ emissions to increase is very high given population growth and worldwide increases in standard of living.

Risk The product of likelihood and consequence of an event or process. For example, the risk of climate change is very high because both the likelihood and the consequences are high.

Storage Resource (capacity) Physical pore-space volume available for CO₂ storage irrespective of economics or regulations.

Storage Reserve (capacity) Pore-space volume available for CO₂ storage including reductions accounting for economic, legal, environmental, and regulatory factors.

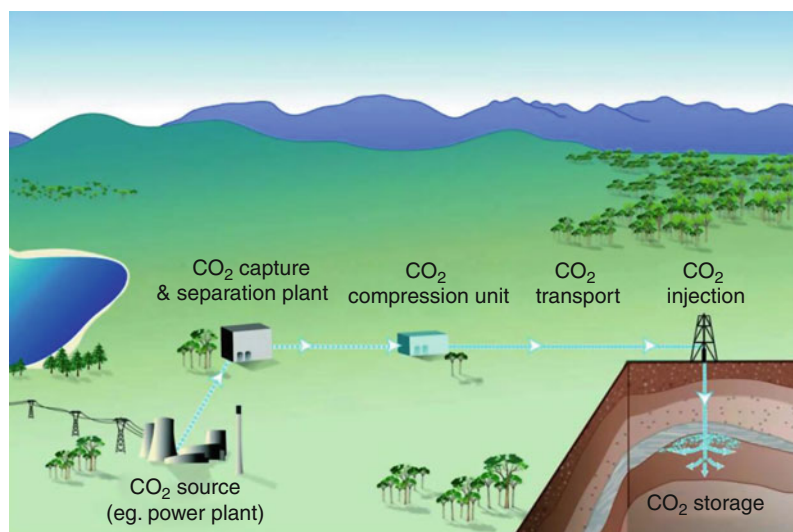
Definition of Subject and Its Importance

Carbon dioxide (CO₂) capture and storage (CCS) is a combination of technologies that addresses climate change by directly reducing the net CO₂ emissions

arising from the use of fossil fuels as the main global primary energy source [1]. In CCS as commonly envisioned, CO₂ will be captured from flue gases at point sources such as coal-fired power plants, compressed, transported by pipeline, and injected into deep geologic formations for permanent storage (i.e., geologic sequestration) (Fig. 1).

The capture of CO₂ involves use of liquid sorbents, membranes, or other advanced materials that can extract CO₂ from a mixture of gases associated with power generation or other industrial processes in which CO₂ is often a minor component at relatively low pressure. Few such capture operations exist currently at more than the pilot scale. However, CO₂ occurs in many natural gas (methane, CH₄) fields at concentrations above those required for delivery to customers. Gas processing to remove CO₂ from natural gas has been carried out for decades, and at two gas fields nearly 2 million tons of captured CO₂ are reinjected annually for geologic sequestration [2, 3].

In power-plant capture, extraction of CO₂ can be done after combustion, so-called post-combustion capture, or during precombustion steps, which has the advantage of higher pressures and higher CO₂ concentrations [1]. There are also direct capture



Geologic Carbon Sequestration: Sustainability and Environmental Risk. Figure 1

Schematic of the carbon dioxide capture and storage (CCS) process (CO₂CRC, <http://www.co2crc.com.au/aboutccs/>)

approaches for CCS that make use of solid sorbents to capture CO₂ from ambient air rather than specifically at point sources [e.g., 4, 5]. Direct-air capture has the advantage that it addresses emissions from all sources, including mobile CO₂ sources such as automobiles and trucks, but also the disadvantages of much lower CO₂ concentration and pressure.

Regardless of how CO₂ capture is accomplished, the process must be capable of providing a stream of CO₂ for compression and transport to sequestration sites. Although direct injection of CO₂ into the deep oceans has received a large amount of attention [e.g., 6], and numerous processes to accelerate uptake of atmospheric CO₂ by the oceans have been discussed [e.g., 7–9], concerns about permanence and impact to marine ecosystems are larger for ocean sequestration than for geologic sequestration [e.g., 10]. This leaves geologic sequestration as the main approach under consideration for isolating from the atmosphere the vast quantities of CO₂ that will need to be captured and stored for CCS to play a role in mitigating climate change.

The extra expense involved in capturing, transporting, and injecting CO₂ in the CCS process can be expressed in terms of an energy penalty, i.e., the amount of energy that must be expended above business-as-usual fossil fuel energy use. Estimates of the energy penalty for CCS vary over a wide range depending on combustion process, age of facility, distance to geologic storage site, etc., but are likely around 40% [1, 11]. Whether stated in terms of dollars or energy penalty, the largest expense in CCS is capture (which also includes compression), currently projected to account for more than 60% of the cost of CCS [12].

Introduction

Fossil fuels are abundant, inexpensive to produce, and are easily converted to usable energy by combustion as demonstrated by mankind's dependence on fossil fuels for over 80% of its primary energy supply [13]. This reliance on fossil fuels comes with the cost of carbon dioxide (CO₂) emissions that exceed the rate at which CO₂ can be absorbed by terrestrial and oceanic systems worldwide resulting in increases in atmospheric CO₂ concentration as recorded by direct

measurements over more than five decades [14]. Carbon dioxide is the main greenhouse gas linked to global warming and associated climate change, the impacts of which are currently being observed around the world, and projections of which include alarming consequences such as water and food shortages, sea-level rise, and social disruptions associated with resource scarcity [15]. The current situation of a world that derives the bulk of its energy from fossil fuel in a manner that directly causes climate change equates to an energy–climate crisis.

Although governments around the world have only recently begun to consider policies to avoid the direst projections of climate change and its impacts, sustainable approaches to addressing the crisis are available. The common thread of feasible strategies to the energy–climate crisis is the simultaneous use of multiple approaches based on available technologies [e.g., 16]. Efficiency improvements (e.g., in building energy use), increased use of natural gas relative to coal, and increased development of renewables such as solar, wind, and geothermal, along with nuclear energy, are all available options that will reduce net CO₂ emissions. While improvements in efficiency can be made rapidly and will pay for themselves, the slower pace of change and greater monetary costs associated with increased use of renewables and nuclear energy suggests an additional approach is needed to help bridge the time period between the present and a future when low-carbon energy is considered cheap enough to replace fossil fuels. Carbon dioxide capture and storage (CCS) is one such bridging technology [1].

CCS has been the focus of an increasing amount of research over the last 15–20 years and is the subject of a comprehensive IPCC report that thoroughly covers the subject [1]. CCS is currently being carried out in several countries around the world in conjunction with natural gas extraction [e.g., 2, 3] and enhanced oil recovery [17]. Despite this progress, widespread deployment of CCS remains the subject of research and future plans rather than present action on the scale needed to mitigate emissions from the perspective of climate change. The reasons for delay in deploying CCS more widely are concerns about cost [18], regulatory and legal uncertainty [19], and potential environmental impacts [21].

This entry discusses the long-term (decadal) sustainability and environmental hazards associated with the geologic CO₂ storage (GCS) component of large-scale CCS [e.g., 20]. Discussion here barely touches on capture and transport of CO₂ which will occur above ground and which are similar to existing engineering, chemical processing, and pipeline transport activities and are therefore easier to evaluate with respect to risk assessment and feasibility. The focus of this entry is on the more uncertain part of CCS, namely geologic storage. The primary concern for sustainability of GCS is whether there is sufficient capacity in sedimentary basins worldwide to contain the large of amounts of CO₂ needed to address climate change. But there is also a link between sustainability and environmental impacts. Specifically, if GCS is found to cause unacceptable impacts that are considered worse than its climate change mitigation benefits, the approach will not be widely adopted. Hence, GCS has elements of sustainability insofar as capacity of the subsurface for CO₂ is concerned, and also in terms of whether the associated environmental risks are acceptable or not to the public.

Carbon Dioxide (CO₂) Capture and Storage (CCS): How Does it Work?

In order to understand the main environmental hazards and sustainability issues associated with GCS, the basic principles of CCS must be understood. First, CO₂ gas compresses into a relatively high-density form at the pressures and temperatures encountered below approximately 1 km in the Earth's crust. In this dense form, called its *supercritical* form because it is neither strictly liquid nor strictly gas, a larger amount of CO₂ can be stored per unit volume than if CO₂ is stored as a gas at shallower levels in the crust. The density of CO₂ at depths greater than 1 km in the crust ranges from around 600 to 850 kg/m³ depending on the geothermal gradient. The maximum depths targeted for GCS are typically in the range of 1–4 km, with the maximum depth dictated by the economics of deep wells and the decreasing permeability of deep sedimentary rock. The density of CO₂ is nearly constant at these depths as the effects on CO₂ density of increasing temperature approximately

compensate for increasing pressure in typical sedimentary basins [21]. Although CO₂ is very dense at depth relative to its gaseous form at the ground surface and can therefore be volumetrically sequestered very efficiently in the deep subsurface, it will always be buoyant relative to the native fluids (saline groundwater or brine) in the subsurface and tend to rise up through them if a flow path is available.

Second, global tectonics have created sedimentary basins on all of the continents in which sediment deposition over geologic timescales has produced thick sequences of sedimentary rock capable of storing CO₂ [22]. There is a vast amount of pore space in these sedimentary rocks arising from the imperfect packing of individual rock grains and incomplete cement filling of the space (pores) between the grains. Significant space can also sometimes arise from pervasive fracturing of the rock. In addition, sedimentary rocks commonly exist in alternating sequences of sandstones (relatively coarse-grained, with high porosity and permeability) and shales (fine-grained, with low permeability) making a configuration in which some sedimentary layers are porous and permeable and others are relatively impermeable. The fine-grained and low-permeability formations are the cap rocks that provide the upper seal for the high-porosity and permeability reservoirs into which CO₂ can be injected in the process of GCS.

Four different primary trapping mechanisms are recognized to occur in the deep subsurface to permanently sequester CO₂ [20]. These include:

1. Structural and stratigraphic trapping, which occurs when buoyant CO₂ flows up and becomes trapped against fine-grained and very low-permeability overlying cap rock often in dome-shaped structures. This same mechanism traps oil and natural gas.
2. Residual gas trapping, the process in which CO₂ bubbles are left trapped in the pores of the rock as CO₂ and water flows through the reservoir (e.g., by buoyancy forces) and water in-fills the pores previously occupied by CO₂. This is the same process that occurs in oil reservoirs as water replaces oil and prevents full recovery motivating various enhanced oil recovery approaches.

3. Solubility trapping, the process in which CO₂ dissolves into the saline water or brine in the reservoir rock. This same process of CO₂ dissolution occurs to create both natural and man-made carbonated beverages.
4. Mineral trapping, which occurs as CO₂ dissolved in the native water reacts with minerals and other dissolved constituents to form new carbonate minerals. This is analogous to the precipitation of travertine that forms in some hot (and cold) spring waters after discharge.

CO₂ injected into the deep subsurface will tend to be trapped by all four of these mechanisms in proportions that vary over time. For example, mineral trapping depends on dissolution [e.g., 23] and precipitation of mineral phases that can take on the order of hundreds to thousands of years [24, 25]. Considered together, the fractions of trapping by residual gas and solubility and mineral precipitation processes tend to increase over time, while the fraction of CO₂ trapped by structural and stratigraphic trapping decreases [20]. As sequestered CO₂ progresses over time through the sequence of structural and stratigraphic, residual gas, solubility, and mineral trapping mechanisms, CO₂ storage is considered to become more permanent over time [20].

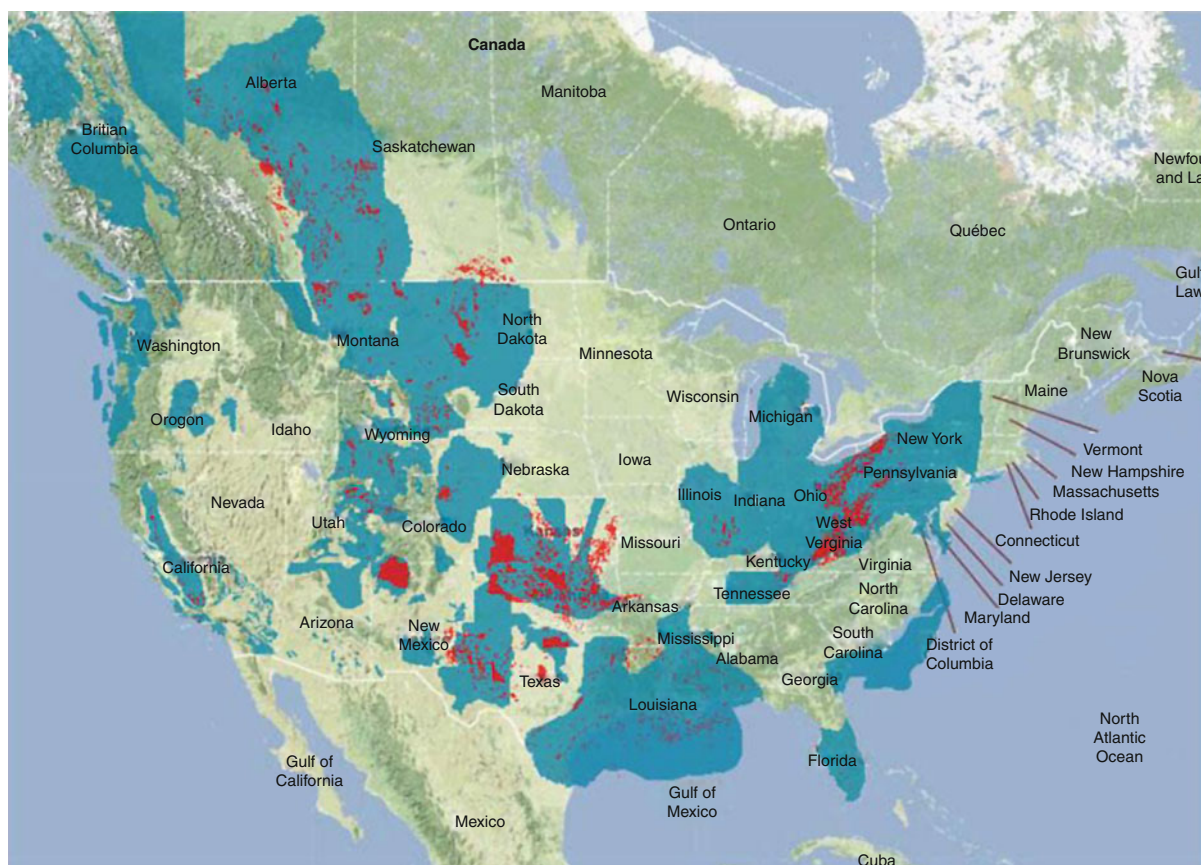
One process that has similarities to GCS is natural gas storage, carried out at over 450 sites in the USA [26]. In this process, methane (CH₄) produced from natural gas reservoirs in one location is reinjected into depleted natural gas reservoirs or aquifer storage reservoirs for temporary storage until market demand (e.g., a cold or hot spell) exceeds supply at which time extra gas is produced from the storage reservoir. In the USA, the amount of natural gas stored is much smaller than the amount of CO₂ that is produced from fossil-fuel power plants (approximately six times less CH₄ by volume (7.5 Tcf = 1.4×10^8 t [27]) is stored overall than there is CO₂ produced at fossil-fuel power plants (47 Tcf = 2.4×10^9 t) per year). Furthermore, the natural gas storage industry uses the same reservoir for decades of injection and production cycles, whereas the GCS industry would need to continuously develop new reservoirs. So while the processes are very similar and much can be learned from

the natural gas storage industry, the scale of the GCS industry will need to be much larger [e.g., 28, 29] in order for it to have an effect on climate change mitigation.

Opportunity and Capacity

As mentioned above, sedimentary basins in the USA and around the world are the primary targets for large-scale GCS [1, 22]. Shown in Fig. 2 are sedimentary basins (blue) in the USA and Canada with hydrocarbon-producing regions shown in red. As shown, there are large areas of the USA and Canada that are potential sinks for CO₂. Most of the opportunity is in sedimentary basins on the continent, but offshore opportunities are also being pursued [e.g., 30]. Economics and regulatory and environmental considerations will govern the extent to which offshore options are viable. Current efforts in North America are mostly aimed at onshore GCS opportunities, while in Europe primarily offshore opportunities are pursued.

Sustainability and feasibility of GCS are largely dependent on capacity. Evaluations have shown there is more than enough capacity to store point-source CO₂ emissions for hundreds of years or more [e.g., 1]. However, large capacity is a necessary but not sufficient condition for GCS feasibility. First, large capacity does not equate to adequate injectivity, i.e., there may be large porosity in some formations that have low permeability or are highly compartmentalized which would require more wells to inject CO₂ than the economics of a project could support. Second, capacity may not be available in close proximity to large CO₂ sources necessitating long pipeline transport distances and associated extra costs [32]. Some of this transport cost can be accommodated under reasonable projections of CO₂ storage economics, but clearly the closer the sink is to the source, the better. Finally, this discussion points out that there are two different types of capacity, namely, resource and reserve capacity [e.g., 33]. Most evaluations to date have focused on resource capacity, i.e., the total amount of pore space available regardless of where it is located or what it takes to access it. As described in the glossary, reserve capacity is the more practical measure of capacity, because it includes not only economics



Geologic Carbon Sequestration: Sustainability and Environmental Risk. Figure 2

Sedimentary basins (*blue*) in the USA and Canada considered good targets for potential geologic storage of CO₂, with oil and gas producing regions shown in *red* [31]

but also policy, environmental, and regulatory restrictions and limitations on capacity. By this definition, reserve capacity is a fraction of resource capacity, and reserve capacity can change over time as economics or regulations change.

Although different methods to estimate capacity have led to wide variations in capacity estimates over various regions [e.g., 34], there is no doubt that there is an enormous amount of resource capacity available. In short, resource capacity does not at present appear to limit the long-term (decadal) sustainability of GCS. On the other hand, resource capacity is not the only measure of feasibility. One must take into account potential environmental impacts associated with GCS such as the possibility of groundwater contamination and induced

seismicity, since the environmental risks and costs may be unacceptable to the public.

Potential Impacts

The injection of large quantities of CO₂ into the deep subsurface through wells is clearly a large perturbation to the local natural system in terms of changing the composition and pressure of the native fluids. Specifically, CO₂ will partially dissolve into the native saline groundwater or brine while also pushing these native fluids outward away from the well as a relatively fast-moving pressure wave. The deep fluid injection process is very well known and practiced widely for injection of various fluids today [19, 26, 35] – and the reverse, production of

fluids through wells such as oil, gas, and groundwater are similarly practiced widely under regulatory frameworks aimed at protecting against adverse consequences. Nevertheless, the novelty of CCS associated with the large volume of CO₂ that needs to be injected motivates discussion of what can go wrong and what general impacts are possible. This discussion will serve to evaluate which impacts are the most likely and which have the greatest consequences. This in turn will allow focus to be placed on the highest environmental risks so they can be avoided altogether, or assessed and mitigated if unavoidable.

Broadly, impacts of CCS can be broken down into those occurring at depth with no discharge of CO₂ into the atmosphere (i.e., the CO₂ storage objective is achieved even as other consequences occur), and those that involve CO₂ discharging into the atmosphere. Presented in Table 1 are potential impacts of geologic CO₂ storage broken down into these two broad categories.

While the impacts arising from CO₂ leaking upward into the vadose zone, root zone, surface water, and out

of the ground may be very serious, such occurrences all require a conduit or flow pathway from the deep injection zone to the near-surface environment, such as an improperly abandoned well or conductive fault. Any GCS project that had moderate to high potential for the leakage scenarios in the upper part of Table 1 would presumably not be undertaken assuming effective risk management, insurance, and regulatory processes are in place. Furthermore, theoretical studies aimed at finding ways that CO₂ could be catastrophically released from CO₂ storage sites leading to the most serious impacts at the ground surface have found self-limiting fluid interference behaviors rather than runaway behaviors [50]. Finally, the impacts described in the upper part of Table 1 are associated with failures of GCS in that CO₂ will enter the atmosphere negating the sequestration objective. Assuming an adequate monitoring program is in place, these leakage events would be relatively obvious and appropriate changes in operations and remedial actions could be carried out.

Geologic Carbon Sequestration: Sustainability and Environmental Risk. Table 1 Shallow (*upper part of table*) and deep (*lower part of table*) processes and potential impacts of geologic CO₂ storage (GCS) (Oldenburg, 2007)

Category	Scenario	Significance	References
CO ₂ enters the atmosphere	Root zone impacts	Profound, visible impact on plants, trees, crops	[36, 37]
	Migration in to vadose zone	May include root zone and entry into buildings	[38]
	Bubbling through surface water	Alters water quality (e.g., lowering pH)	[39]
	Accumulation in topographic lows	Very hazardous due to possibility of asphyxiation	[38, 40]
	Seepage into basements and homes	Very hazardous due to possibility of asphyxiation	[41]
	Ground plumes	Very hazardous due to possibility of asphyxiation	[40, 42, 43]
CO ₂ may or may not enter atmosphere	Intrusion of CO ₂ into potable water	Lowers pH, dissolves minerals potentially releasing heavy metals	[44, 45]
	Intrusion of CO ₂ into hydrocarbon, mineral, or geothermal resources	Lowers value of natural gas or mineral resources such as potash	
	Displacement of saline groundwater or brine into potable water by regional pressurization	Saline water intrusion into potable water degrades water quality	[45–47]
	Induced seismicity	CO ₂ injection pressure may cause felt earthquakes	[48, 49]

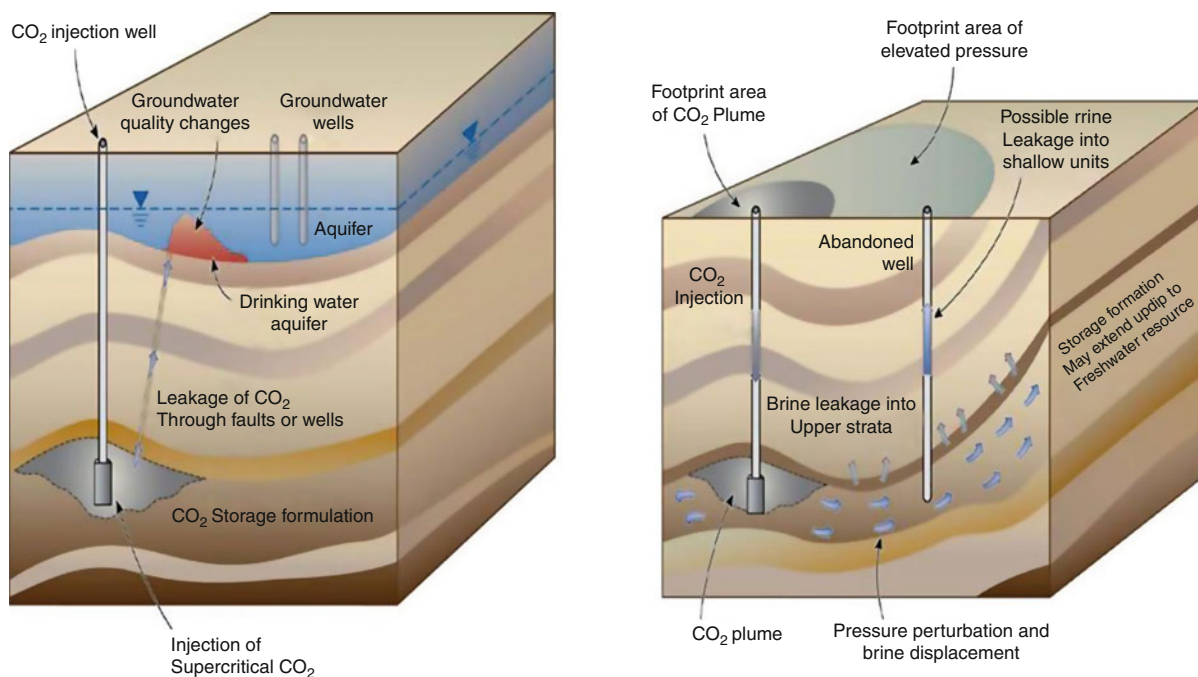
In the exceptional case of the occurrence of an uncontrolled CO₂ leak from a well into the atmosphere, the main consequence of concern is asphyxiation of workers or bystanders. Documented CO₂ well blowouts associated with oil production indicate the asphyxiation hazard is low for blowouts occurring in open environments [e.g., 51, 52]. Modeling studies of open-air scenarios have also found that the area of asphyxiation hazard around a blowout is small because dispersion acts to rapidly decrease concentrations [53].

In contrast to a well blowout or the scenarios in the upper part of Table 1, it may be much more difficult to detect the onset and development of the scenarios listed in the lower part of the table in order to take early action to limit impacts. Although the scenarios listed in the lower part of Table 1 do not involve CO₂ entering the atmosphere and thus do not involve outright failure of GCS, the intrusion of CO₂ or saline groundwater or brine into groundwater resources and injection-induced seismicity are considered the main hazards associated with GCS that are likely enough to warrant risk assessment

and related regulatory measures in order to minimize the likelihood of their occurrence and their consequences. These two categories of risk, described in more detail below, must be assessed and managed as part of widespread, long-term, and sustainable GCS deployment.

Potential Impacts to Potable Groundwater

CO₂ that leaks upward out of the storage region through wells [e.g., 54, 55], or faults and fractures [56], can potentially enter potable groundwater resources as shown schematically in Fig. 3a. Degradation of the groundwater quality is possible through indirect contamination. As CO₂ dissolves into groundwater, it partitions into species comprising dissolved inorganic carbon (DIC) as CO₂(aq), HCO₃⁻, and CO₃²⁻, resulting in a decrease in the solution pH. At the same time, alkalinity is controlled by HCO₃⁻ and CO₃²⁻, which can increase upon CO₂ dissolution. Control over the geochemical changes in the water is provided by the composition and mineralogy of the mineral grains, coatings, and cements present in the rock. For example,



Geologic Carbon Sequestration: Sustainability and Environmental Risk. Figure 3

Potential groundwater impact scenarios. Left-hand figure from [45], right-hand figure from [47]

a carbonate mineral in the rock such as calcite (CaCO_3) will dissolve by the reaction $\text{CO}_2 + \text{H}_2\text{O} + \text{CaCO}_3 = \text{Ca}^{2+} + 2\text{HCO}_3^-$, resulting in the doubling of dissolved inorganic carbon (DIC) (i.e., one mole of CO_2 reacts to produce two moles of HCO_3^-) and a release of Ca^{2+} to solution. Similar reactions are possible involving alteration of biotite, plagioclase, and alkali feldspar, and other common minerals in sedimentary rocks [e.g., 25].

CO_2 or saline groundwater and brine leakage into groundwater aquifers will also give rise to impacts on microbiological communities [57]. Although cell density declines by three to six orders of magnitude from the ground surface to 4 km depth, microbes at the depths of potable groundwater can be affected if CO_2 or brine intrudes into this region. The alteration of minerals such as feldspars by acidic groundwater can release iron which can stimulate Fe^{3+} -reducing communities and result in methanogenesis. Clearly, microbial processes can affect geochemistry and vice versa.

Assuming the reaction kinetics allow it, geochemical reactions can further alter pH, DIC, isotopic composition, and trace element concentrations in solution. For example, trace elements in the minerals, in coatings, or in ion exchange sites in clays (including heavy metals such as lead) may be released into groundwater as biogeochemical conditions change with associated degradation of groundwater quality [44, 58]. Observations of such effects have been made during CO_2 injection experiments at field sites [e.g., 59, 60] and in the laboratory [61]. Recent work has further assessed the potential for such reactions by examining actual groundwater compositions and aquifer mineralogy from across the USA, and found that increases in the concentration of As and Pb could be a concern if widespread CO_2 leakage into groundwater resources were to occur [45]. Buffering reactions may serve to moderate pH decline and may serve to diminish groundwater degradation as observed in a natural analog study in New Mexico [62]. In summary, it is recognized that impacts of CO_2 leakage on potable groundwater may be significant and costly if they occur, and therefore careful monitoring, GCS operations, and site selection [e.g., 63] are essential to reduce groundwater contamination risk.

Another hazard to groundwater resources is the potential intrusion of displaced saline groundwater or brine or CO_2 -charged water into potable groundwater as shown in Fig. 3b. In addition to the above

biogeochemical impacts arising from the CO_2 itself, there is the first-order degradation arising from the presence of dissolved solids (e.g., NaCl, CaCl_2 , and KCl) in the saline groundwater or brine along with whatever trace elements it may contain. Potable groundwater in the USA is defined on the basis of total dissolved solids (TDS) content equal to 10,000 mg/L or less. Injection into deep aquifers is regulated in the USA by the Environmental Protection Agency (EPA) under the Underground Injection Control (UIC) program to protect potable groundwater from degradation [e.g., 35, 64]. The hazard arising from GCS is that deep saline water or brine pressurized by CO_2 injection may tend to migrate upward into potable groundwater aquifers, thereby increasing TDS and degrading the resource.

The main reason that saline groundwater or brine intrusion arising from GCS is such a concern is that pressure increases associated with CO_2 injection can occur at great distances (~10–100 km) from the injection site [46, 47, 65, 66]. So while characterization of a given site may have demonstrated that CO_2 will be contained within a well-defined CO_2 storage region, there will generally be a large region of pressure increase in the formation that may not have been characterized to the same degree because of the large distance from the injection site. Because of this, it is possible that the cap rock may not be continuous over these large distances, or may not have the same integrity as the region targeted for CO_2 storage. Nevertheless, in order for upward saline groundwater or brine intrusion to occur, there must be a driving force in addition to the conduit or pathway (e.g., improperly abandoned well, or fault or fracture zone). Although pressure rise is high near the CO_2 injection wells, it falls off rapidly away from the injection wells. In addition, dense brines with high TDS require overpressures to be driven upward into potable groundwater through wells or other conduit (e.g., conductive fault) due to their high density and resistance to flow [67]. Furthermore, once in the potable aquifer, the higher density of the brine will tend to limit the extent of its mixing with potable groundwater [68].

Induced Seismicity

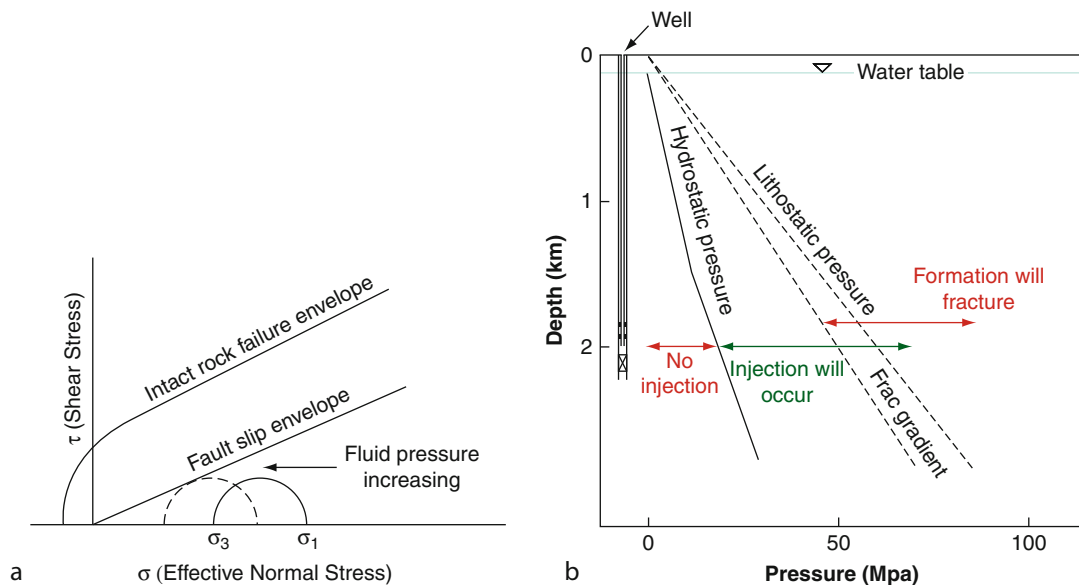
The phenomenon of induced seismicity due to fluid injection has been recognized for approximately 50 years starting with the well-known example of injected fluid waste disposal at the Rocky Mountain

Arsenal in Colorado [69–71]. Induced seismicity is well understood from experience in the fields of injection for deep disposal of liquid waste, and injection for geothermal energy extraction [48, 49, 72]. Induced seismicity is caused by the reduction in effective stress that accompanies an increase in pore pressure. The potential for induced seismicity is determined by the Mohr–Coulomb criterion which quantifies the amount of normal stress beyond that provided by fluid pressure that is needed before shear failure occurs (i.e., reactivation of existing faults or slippage along fractures). The Mohr–Coulomb criterion is given by the relation $\tau = C + \mu (\sigma_n - p)$ where τ is the shear strength of the rock, C is the Coulomb criterion, μ is the coefficient of internal friction, σ_n is the normal stress, and p is the fluid pressure [e.g., 73]. When the right-hand side (normal stress) is smaller than the left-hand side (shear strength), the rock is likely to slip along fracture planes of optimal orientation, which releases seismic energy (i.e., causes an earthquake). The Mohr–Coulomb equation shows that injection pressure reduces the effective normal stress in the rock, hence the tendency for injection to cause slippage

along existing faults and fractures as shown in Fig. 4a.

A simple graphical representation of pressures as a function of depth helps elucidate the processes active near an injection well. Shown in Fig. 4b are the variations with depth of hydrostatic pressure, so-called fracture pressure (or commonly frac pressure), and lithostatic pressure. As shown, fluid pressure at an injection well must be larger than hydrostatic pressure in order for injection to occur. However, if the pressure exceeds the frac pressure at a given depth, the injection process will tend to fracture the formation. By the Mohr–Coulomb criterion, seismicity can be induced at injection pressures below the frac pressure as effective stress decreases and existing faults are reactivated. Either the generation of new fractures, or slippage along existing faults and fractures, is manifest as induced seismicity.

It is important to note that while the word *earthquake* evokes fear and a certain image of destruction in most people's minds, the term encompasses a wide range of magnitudes, from microseismic earthquakes



Geologic Carbon Sequestration: Sustainability and Environmental Risk. Figure 4

(a) Mohr circle representation of fault slip (induced seismicity) as fluid pressure increases, and (b) pressure-depth depiction showing hydrostatic, frac, and lithostatic pressure gradients

that cannot be felt by humans, to great earthquakes that imperil life and damage structures. Earthquakes tend to follow a logarithmic frequency distribution such that very small earthquakes are orders of magnitude more frequent than very large earthquakes [74]. Experience from water injection into geothermal systems shows that the majority of induced seismicity is microseismicity, with felt earthquakes much rarer, and moderate to large earthquakes rarer still [48]. Despite the fact that large earthquakes are not expected to be induced by CO₂ injection in carefully chosen sites [75], the hazard of induced seismicity is looming large at present in the area of public acceptance of GCS.

Aside from the hazard of ground acceleration at the surface, induced seismicity also creates the possibility that a cap rock seal could fracture or a fault could become permeable giving rise to a leakage pathway for CO₂ [e.g., 73, 76]. This is a well-recognized failure mode, and injection regulations are aimed at preventing fracturing from happening. However, induced seismicity of critically stressed rocks on preexisting faults and fractures is still possible even when the frac pressure is not exceeded. The extent to which the risk of induced seismicity, objectively considered to be a small risk, outweighs the benefits in terms of risk reduction of climate change that CCS affords, is one of the questions that must be addressed by the public and decision-makers to guide their acceptance of CCS.

Future Directions

As the discussion here suggests, one approach that can aid in addressing the energy–climate crisis is CCS. There are significant costs to CCS, primarily associated with capture, and CCS also brings with it recognized environmental risks the most uncertain of which are associated with the geologic storage component of the process. The main risks in GCS are threat to potable groundwater and induced seismicity, two areas of active research. Despite the need for greater understanding of these hazards, mitigation measures are available today. For example, if contamination of groundwater were to occur, the water could be treated, or alternate sources could be found if treatment is found impractical [63]. As for induced seismicity, the hazard can be reduced by reducing injection pressure

(e.g., through use of more wells for a given CO₂ source), carrying out pressure management through saline groundwater or brine extraction, by careful site selection that avoids heavily faulted and tectonically active areas, and by establishing and enforcing building codes.

The path forward for demonstrating and deploying CCS, as a sustainable part of the portfolio of energy production and use changes that are needed to mitigate the energy–climate crisis, can be described as follows. First, testing and demonstration projects [e.g., 77, 78] that include CO₂ capture from anthropogenic sources need to expand rapidly and by many factors so that the technology can be perfected in different regions and geologic settings. These multiple demonstration projects will show how GCS works, and if GCS continues to perform as envisioned, additional CCS deployments can be added over time. Second, research on capture and alternative combustion approaches that enable more efficient capture should be accelerated. Third, a large program of site characterization and capacity studies [e.g., 79] should be undertaken so that the well-known large basin-scale sites are understood and operational plans can be put in place quickly at the time when large-scale capture facilities come on line and anthropogenic CO₂ streams become available for sequestration. Fourth, research on injection, trapping, migration, long-term fate, leakage impacts, mitigation, monitoring, and modeling should be continued so that GCS can be optimized and related technologies can be commercialized and deployed in a cost-effective manner. Fifth, governments at all levels should promulgate regulatory and economic policies that answer the current questions and uncertainty faced by businesses who foresee the broad outlines of a carbon-constrained future but do not yet have the clear ground rules (e.g., policies on carbon pricing, injection regulations, and legal frameworks) provided by government that are necessary for making the large capital investments required for CCS.

The decision to take on the costs and risks of GCS, with the accompanying promise of contributing to reductions in the extent of climate change, should be made based on an objective comparison against the climate and environmental risks of carrying on business as usual with fossil-fuel use and unabated CO₂ emissions. The public and decision-makers should

keep in mind that the environmental risks of CCS are local to the basin where GCS is carried out, whereas the projected impacts of climate change are global-to-regional in scale and are expected to have profound consequences for the social, physical, and natural systems on Earth. Support for GCS technology will come in the form of policy decisions about carbon pricing, injection regulations, and legal frameworks that encourage commercial applications of CCS. The decision about whether to adopt these policies will ultimately fall on the public or its representatives. The risks to the Earth's environment and social systems of doing nothing about the energy-climate crisis must be communicated effectively to the public and the decision-makers so that they can make an informed decision about acceptable risks and costs of the various options available for avoiding the worst impacts of climate change.

Acknowledgments

This entry greatly benefitted from suggestions and comments by my LBNL colleagues Karsten Pruess, Jens Birkholzer, and Preston Jordan.

Bibliography

Primary Literature

- Metz B, Davidson O, de Coninck HC, Loos M, Meyer LA (eds) (2005) IPCC special report on carbon dioxide capture and storage. Prepared by working group III of the intergovernmental panel on climate change. Cambridge University Press, Cambridge
- Torp TA, Gale J (2004) Demonstrating storage of CO₂ in geological reservoirs: the Sleipner and SACS projects. *Energy* 29(9–10):1361–1369
- Mathieson A, Midgley J, Dodds K, Wright I, Ringrose P, Saoul N (2010) CO₂ sequestration monitoring and verification technologies applied at Krechba, Algeria. *Lead Edge* 29(2):216–222
- Keith DW, Ha-Duong M, Stolaroff JK (2006) Climate strategy with CO₂ capture from the air. *Clim Change* 74(1–3):17–45
- Lackner K (2010) Washing carbon out of the air. *Sci Am Mag Sci Am* 302:66–71. doi:10.1038/scientificamerican0610-66
- Brewer PG, Friederich G, Peltzer ET, Orr FM Jr (1999) Direct experiments on the ocean disposal of fossil fuel CO₂. *Science* 284(5416):943–945
- Caldeira K, Rau GH (2000) Accelerating carbonate dissolution to sequester carbon dioxide in the ocean: geochemical implications. *Geophys Res Lett* 27(2):225–228
- Chisholm SW, Falkowski PG, Cullen JJ (2001) Dis-crediting ocean fertilization. *Science* 294(5541):309–310
- Buesseler KO, Boyd PW (2003) Will ocean fertilization work? *Science* 300(5616):67–68
- Shaffer G (2010) Long-term effectiveness and consequences of carbon dioxide sequestration. *Nat Geosci* 3:464–467
- House KZ, Harvey CF, Aziz MJ, Schrag DP (2009) The energy penalty of post-combustion CO₂ capture & storage and its implications for retrofitting the U.S. installed base. *Energy Environ Sci*. doi:10.1039/b811608c
- McKinsey and Co. (2008) Carbon capture and storage: assessing the economics. McKinsey&Company, New York, NY, p 49
- IEA, Key World Energy Statistics (2009) International Energy Agency (IEA), Paris, http://www.iea.org/textbase/nppdf/free/2009/key_stats_2009.pdf
- Keeling RF, Piper SC, Bollenbacher AF, Walker JS (2009) In: Trends A (ed) Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center. Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn. doi:10.3334/CDIAC/atg.035 (Atmospheric CO₂ records from sites in the SIO air sampling network)
- Core Writing Team, Pachauri, RK and Reisinger, A (eds) (2007) IPCC, Climate change 2007: synthesis report. Contribution of working groups I, II and III to the fourth assessment report of the intergovernmental panel on climate change. IPCC, Geneva, p 104
- Pacala S, Socolow R (2004) Stabilization wedges: solving the climate problem for the next 50 years with current technologies. *Science* 305(5686):968–972
- Preston C, Monea M, Jazrawi W, Brown K, Whittaker S, White D, Law D, Chalaturnyk R, Rostron B (2005) IEA GHG Weyburn CO₂ monitoring and storage project. *Fuel Process Technol* 86(14–15):1547–1568
- Rubin ES, Chen C, Rao AB (2007) Cost and performance of fossil fuel power plants with CO₂ capture and storage. *Energy Policy* 35(9):4444–4454
- Wilson EJ, Gerard D (2007) Risk assessment and management for geologic sequestration of carbon dioxide. In: Wilson EJ, Gerard D (eds) Carbon capture and sequestration integrating technology, monitoring, and regulation. Blackwell, Oxford, pp 101–125
- Benson SM, Cook PJ (eds) (2005) IPCC Special report on CO₂ capture and storage, chap. 5. Cambridge University Press, Cambridge
- Oldenburg CM (2007) Migration mechanisms and potential impacts of CO₂ leakage and seepage. In: Wilson E, Gerard D (eds) Carbon capture and sequestration integrating technology, monitoring, and regulation. Blackwell, Ames, pp 127–146 (BNL-58872)
- Bachu S (2003) Screening and ranking of sedimentary basins for sequestration of CO₂ in geological media in response to climate change. *Environ Geol* 44(3):277–289

23. Spycher N, Ennis-King J, Pruess K (2003) CO₂-H₂O mixtures in the geological sequestration of CO₂. Assessment and calculation of mutual solubilities from 12 C to 100 C and up to 600 bar. *Geochemica Cosmochim Acta* 67:3015–3031
24. Gunter WD, Bachu S, Benson S (2004) The role of hydrogeological and geochemical trapping in sedimentary basins for secure geological storage of carbon dioxide. *Geol Soc Lond Spec Publ* 233:129–145
25. Xu T, Apps JA, Pruess K (2005) Mineral sequestration of CO₂ in a sandstone-shale system. *Chem Geol* 217 (1–4):295–318
26. Benson SM, Hepple R, Apps J, Tsang C-F, Lippmann M (2002) Lessons learned from natural and industrial analogues for storage of carbon dioxide in deep geological formations. E.O. Lawrence Berkeley National Laboratory Report LBNL-51170
27. Katz DL, Tek MR (1981) Overview on underground storage of natural gas. *J Petrol Technol* 33(6):943–951
28. USGS (2010) http://energy.er.usgs.gov/health_environment/co2_sequestration/co2_illustrations.html. Accessed 6 Oct 2010
29. Haszeldine RS (2009) Carbon capture and storage: how green can black be? *Science* 325(5948):1647–1652
30. Shrag DP (2009) Storage of carbon dioxide in offshore sediments. *Science* 325:1658–1659
31. NATCARB (2008) U.S. Department of energy, carbon sequestration Atlas of the United States and Canada, Office of fossil energy, National energy technology laboratory. <http://geoportal.kgs.ku.edu/natcarb/atlas08/gsinks.cfm>. Accessed 6 Oct 2010
32. McCoy ST, Rubin ES (2008) An engineering-economic model of pipeline transport of CO₂ with application to carbon capture and storage. *Int J Greenhouse Gas Control* 2(2):219–229
33. Brennan ST, Burruss RC, Merrill MD, Freeman PA, Ruppert LF (2010) A probabilistic assessment methodology for the evaluation of geologic carbon dioxide storage: U.S. Geological survey open-file report 2010–1127, p 31. Available only at <http://pubs.usgs.gov/of/2010/1127>. Accessed 6 Oct 2010
34. Bradshaw J, Bachu S, Bonijoly D, Burruss R, Holloway S, Christensen NP, Mathiassen OM (2007) CO₂ storage capacity estimation: issues and development of standards. *Int J Greenhouse Gas Control* 1:62–68
35. USEPA (United States Environmental Protection Agency), Technical program overview, underground injection control regulations, office of water 4606, EPA 816-R-02-025. Revised July 2001. <http://www.epa.gov/safewater/uic/index.html>. Accessed 10 Oct 2010
36. Qi J, Marshall JD, Matson KG (1994) High soil carbon dioxide concentrations inhibit root respiration of Douglas Fir. *New Phytology* 128:435–441
37. Farrar CD, Sorey ML, Evans WC, Howie JF, Kerr BD, Kennedy BM, King C-Y, Southon JR (1995) Forest-killing diffuse CO₂ emission at mammoth mountain as a sign of magmatic unrest. *Nature* 376:675–677
38. Oldenburg CM, Unger AJA (2003) On leakage and seepage from geologic carbon sequestration sites: unsaturated zone attenuation. *Vadose Zone J* 2(3):287–296
39. Oldenburg CM, Lewicki, JL (2005) Leakage and seepage of CO₂ from geologic carbon sequestration sites: CO₂ migration into surface water, *Lawrence Berkeley National Laboratory Report LBNL-57768*
40. Giggensbach WF, Sano Y, Schmincke HU (1991) CO₂-rich gases from lakes nyos and monoun, cameroon; laacher see, Germany, Dieng, Indonesia, and Mt. Gambier, Australia-variations on a common theme. *J Volcanol Geotherm Res* 45:311–323
41. Robinson AL, Sextro RG, Riley WJ (1997) Soil-gas entry into houses driven by atmospheric pressure fluctuations—the influence of soil properties. *Atmos Environ* 31(10):1487–1495
42. Hanna SR, Steinberg KW (2001) Overview of Petroleum Environmental Research Forum (PERF) dense gas dispersion modeling project. *Atmos Environ* 35:2223–2229
43. Britter RE (1989) Atmospheric dispersion of dense gases. *Ann Rev Fluid Mech* 21:317–344
44. Wang S, Jaffe PR (2005) Dissolution of a mineral phase in potable aquifers due to CO₂ releases from deep formations; effect of dissolution kinetics. *Energy Convers Manage* 45:2833–2848
45. Apps JA, Zheng L, Zhang Y, Xu T, Birkholzer JT (2010) Evaluation of potential changes in groundwater quality in response to CO₂ leakage from deep geological storage. *Transp Porous Media* 82:215–246
46. Birkholzer JT, Zhou Q, Tsang C-F (2009) Large-scale impact of CO₂ storage in deep saline aquifers: a sensitivity study on the pressure response in stratified systems. *Int J Greenhouse Gas Control* 3(2):181–194
47. Birkholzer JT, Zhou Q (2009) Basin-Scale hydrogeologic impacts of CO₂ storage: capacity and regulatory implications. *Int J Greenhouse Gas Control* 3(6):745–756
48. Majer EL, Baria R, Stark M, Oates S, Bommer J, Smith B, Asanuma H (2007) Induced seismicity associated with enhanced geothermal systems. *Geothermics* 36(3):185–222
49. Majer E, Baria R, Stark M (2008) Protocol for induced seismicity associated with enhanced geothermal systems. Report produced in Task D Annex I (, International energy agency-geothermal implementing, agreement (incorporating comments by: Bromley C, Cumming W, Jelacic A and Rybach A) (<http://www.iea-gia.org/publications.asp>)
50. Pruess K (2005) Numerical studies of fluid leakage from a geologic disposal reservoir for CO₂ show self-limiting feedback between fluid flow and heat transfer. *Geophys Res Lett* 32(14):L14404
51. Skinner L (2003) CO₂ blowouts: an emerging problem. *World Oil* 224(1):38–42
52. Gouveia FJ, Johnson M, Leif, RN, Friedmann SJ (2005) Aerometric measurement and modeling of the mass of CO₂ emissions from Crystal Geyser, Utah. NETL 4th Annual carbon capture and sequestration conference. Alexandria, p 2–5

53. Aines RD, Leach MJ, Weisgraber TH, Simpson MD, Friedman SJ, Bruton CJ (2008) Quantifying the potential exposure hazard due to energetic releases from a failed sequestration well. In: *Proceedings of the ninth international conference on greenhouse gas control technologies GHGT-9. Washington, p 16–20*
54. Gasda SE, Bachu S, Celia MA (2004) Spatial characterization of the location of potentially leaky wells penetrating a deep saline aquifer in a mature sedimentary basin. *Environ Geol* 46:707–720
55. Scherer GW, Celia MA, Prevost J-H, Bachu S, Bruant R, Duguid A, Fuller R, Gasda SE, Radonjic M, Vichit-Vadkan W (2005) Leakage of CO₂ through abandoned wells: role of corrosion of cement. In: Thomas DC, Benson SM (eds) *Carbon dioxide capture for storage in deep geologic formations*, vol 2. Elsevier, Amsterdam, pp 827–848
56. Shipton ZK, Evans JP, Kirschner D, Kolesar PT, Williams AP, Heath J (2004) Analysis of CO₂ leakage through 'low-permeability' faults from natural reservoirs in the Colorado Plateau, southern Utah. In: Baines SJ, Worden RH (eds) *Geological storage of carbon dioxide*. Geological society, vol 233. Special Publications, London, pp 43–58
57. Onstott TC (2005) Impact of CO₂ injections on deep subsurface microbial ecosystems and potential ramifications for the surface biosphere. In: Thomas DC, Benson SM (eds) *Carbon dioxide capture for storage in deep geologic formations*, vol 2. Elsevier, London, pp 1217–1249
58. Schuett H, Wigand M, Spangenberg E (2005) Geophysical and geochemical effects of supercritical CO₂ on sandstones. In: Thomas DC, Benson SM (eds) *Carbon dioxide capture for storage in deep geologic formations*, vol 2. Elsevier, Amsterdam, pp 767–786
59. Kharaka Y, Cole DR, Hovorka SS, Gunther WD, Knauss KG, Freifeld BM (2006) Gas-water-rock interactions in Frio formation following CO₂ injection: implications for the storage of greenhouse gases in sedimentary basins. *Geology* 34:577–580
60. Kharaka YK, Thordsen JJ, Kakouros E, Ambats G, Herkelrath WN, Birkholzer JT, Apps JA, Spycher NF, Zheng L, Trautz RC, Rauch HW, Gullickson K (2010) Changes in the Chemistry of Shallow Groundwater Related to the 2008 Injection of CO₂ at the ZERT Field, SiteBozeman, Montana. *Environ Earth Sci* 60:273–284
61. Carroll S (May 2009) Trace metal release from Frio sandstone reacted with CO₂ and 1.5 N NaCl Brine at 60°C. In: *Proceedings 8th Annual conference on carbon capture and sequestration*. Pittsburgh
62. Keating EH, Fessenden J, Kanjorski N, Koning DJ, Pawar R (2010) The impact of CO₂ on shallow groundwater chemistry: observations at a natural analog site and implications for carbon sequestration. *Environ Earth Sci* 60(3):521–536
63. Price PN, Oldenburg CM (2009) The consequences of failure should be considered in siting geologic carbon sequestration projects. *Int J Greenhouse Gas Control* 3(5):658–663
64. Wilson EJ, Johnson TL, Keith DW (2003) Regulating the Ultimate Sink: managing the risks of geologic CO₂ storage. *Environ Sci Technol* 37(16):3476–3483
65. Nicot J-P (2008) Evaluation of large-scale CO₂ storage on fresh-water sections of aquifers: an example from the Texas Gulf Coast Basin. *Int J Greenhouse Gas Control* 2(4):582–593
66. Zhou Q, Birkholzer JT, Mehnert E, Lin Y-F, Zhang K (2009) Modeling basin- and plume-Scale Processes of CO₂ storage for full-scale deployment. *Ground Water* 48(4):494–514
67. Nicot J-P, Oldenburg CM, Bryant SL, Hovorka SD (2009) Pressure perturbations from geologic carbon sequestration: area-of-review boundaries and borehole leakage driving forces. *Energy Procedia* 1(1):47–54
68. Oldenburg CM, Rinaldi AP (2011) Buoyancy effects on upward brine displacement caused by CO₂ injection, *Transport in Porous Media* 87(2):525–540
69. Hollister JC, Weimer RJ (eds) (1968) *Geophysical and geological studies of the relationships between the Denver earthquakes and the Rocky Mountain Arsenal well*. Q. Colorado School of Mines 63, Golden, p 251
70. Hoover DB, Dietrich JA (1969) Seismic activity during the 1968 test pumping at the Rocky Mountain arsenal disposal well, circular 613, U.S. Geological Survey, Washington
71. Herrmann RB, Park S-K, Wang C-Y (1981) The Denver earthquakes of 1967–1968. *Bull Seismol Soc Am* 71(3):731–745
72. Cypser DA, Davis SD (1998) Induced seismicity and the potential for liability under U.S. law. *Tectonophysics* 289(1–3):239–255
73. Rutqvist J, Birkholzer J, Cappa F, Tsang C-F (2007) Estimating maximum sustainable injection pressure during geological sequestration of CO₂ using coupled fluid flow and geomechanical fault-slip analysis. *Energy Convers Manage* 48(6):1798–1807
74. Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. *Bull Seismol Soc Am* 17:185–188
75. Sminchak J, Gupta N, Byrer C, Bergman P (2003) Aspects of induced seismic activity and deep-well sequestration of carbon dioxide. *Environ Geosci* 10(2):81–89
76. Wiprut D, Zoback M (2000) Fault reactivation and fluid flow along a previously dormant normal fault in the northern North Sea. *Geology* 28(7):595–598
77. Hovorka SD, Benson SM, Doughty C, Freifeld BM, Sakurai S, Daley TM, Kharaka YK, Holtz MH, Trautz RC, Nance HS, Myer LR, Knauss KG (2006) Measuring permanence of CO₂ storage in saline formations: the Frio experiment. *Environ Geosci* 13(2):105–121
78. Litynski J, Plasynski S, McIlvried HG, Mahoney C, Srivastava RD (2008) The United States department of energy's regional carbon sequestration partnerships program validation phase. *Environ Int* 34(1):127–138
79. Friedmann SJ, Dooley JJ, Held H, Edenhof O (2006) The low cost of geological assessment for underground CO₂ storage: policy and economic implications. *Energy Convers Manage* 47(13–14):1894–1901

Books and Reviews

- Baines SJ, Worden RH (eds) (2004) *Geologic storage of carbon dioxide*, geological society, vol 233. Special Publications, London, pp 1–247
- Eide LI (2009) *Carbon dioxide capture for storage in deep geological formations*, vol 3. CPL Press/BP, Newbury, Berkshire
- Thomas DC, Benson SM (eds) (2007) *Carbon dioxide capture for storage in deep geologic formations—results from the CO₂ capture project*, vol 2. Elsevier, Kidlington, Oxford
- Wilson EJ, Gerard D (eds) (2007) *Carbon capture and sequestration integrating technology, monitoring, and regulation*, Blackwell Publishing, Ames, Iowa

Geothermal Conditioning: Critical Sources for Sustainability

NINA BAIRD

Center for Building Performance & Diagnostics,
Carnegie Mellon University, Pittsburgh, PA, USA

Article of Outline

Glossary
Definition of the Subject
Introduction
The Earth's Thermal Energy Sources and Ground Temperature Distribution
Geothermal Conditioning Principles and Approaches
Installed Capacity and Annual Energy Use
Future Directions
Bibliography

Glossary

Closed loop system A continuous, sealed, underground or submerged heat exchanger through which a heat transfer fluid passes to and returns from building conditioning equipment.

Geothermal direct use Use of thermal energy in the earth or earth-coupled fluid as a heat source and heat transfer reservoir for heating or cooling, without further conversion such as electric power generation.

Geothermal heat pump A conditioning device that uses the ground or ground-coupled fluid as the heat source and heat sink in the heat pump's process

of extracting heat from a low-temperature source and transferring it to sink at a higher temperature by adding the work of a refrigerant, usually with a vapor compression-expansion cycle.

Low-Exergy System Heating and/or cooling system that uses energy at a temperature close to room temperature for efficient utilization of low-grade energy sources.

Open Loop System A system designed to use groundwater or surface water for the purpose of extracting or rejecting heat for building conditioning.

Underground Thermal Energy Storage (UTES) A subsurface system for storing heat and/or cold using groundwater and/or the ground in natural or constructed media.

Definition of the Subject

Geothermal conditioning is use of the earth's thermal energy and storage capacity for heating, cooling, and ventilation. These types of conditioning strategies can transfer heat to the indoor environment using the ground, groundwater, or surface water – resources that are abundant and ubiquitous – to satisfy some or all of the heating load. They can also capitalize on the heat capacity and thermal inertia of the earth and its waters by transferring excess heat from indoors to outdoors, providing cooling with substantially reduced energy consumption from conventional cooling and negligible thermal impact on the outdoor environment.

Geothermal conditioning, like solar conditioning, includes passive and active strategies. Both have a long history. Passive earth sheltering has been used by plant and animal species throughout time. Active geothermal conditioning as defined here for heating, cooling, and ventilation was not broadly tracked until 1995 when geothermal heat pumps were added to global reports of geothermal direct use installations [1]. However, frequently cited early examples are the district heating system installed in Boise, Idaho (USA) in 1892 to heat 400 homes [2] and a residential geothermal steam heat system in Tuscany, Italy, introduced between 1910 and 1940 [3]. In present times, the Paleiskwartier district geothermal system in 's Hertogenbosch, the Netherlands, includes a large pond and aquifer thermal energy storage (ATES) coupled with heat pumps to condition a mixed-use development of 1,200 housing

units and more than 135,000 m² of office space, retail, and entertainment [4].

Accessible geothermal energy includes ground temperatures in excess of 150°C, hot enough to generate electricity and higher than temperatures that can be used directly for building conditioning. In such situations, building conditioning can be linked to geothermal power production through energy cascades, but the cost and complexity of power production processes substantially exceed those for building conditioning alone. In addition, access to geothermal resources that support electricity generation is far more limited. For those reasons, this entry will focus on direct use geothermal conditioning. References that address geothermal power production and geothermal energy cascades are provided in the [Future Directions](#) and [Books and Reviews](#) sections.

Introduction

In 1973, an Icelandic engineer, Baldur Lindal, listed current and potential uses of geothermal energy for conditioning and industrial processes and their temperature requirements. That list became known as the Lindal diagram, shown in [Fig. 1](#) with conditioning processes highlighted. Since that time, substantial improvements in the availability and use of insulation products have allowed well-designed building enclosures to manage an increasing portion of the conditioning loads. When that occurs, heating and cooling can be provided at temperatures much closer to the human comfort range, no longer needing to overcome extreme heat loss or gain at the perimeter. Sustainably designed buildings can satisfy occupant comfort requirements with smaller mechanical equipment that operates at lower supply temperatures, supporting low-exergy systems (see [Fig. 2](#)). This puts building conditioning squarely in the geothermal “sweet spot,” a temperature range available to almost every building in contact with the earth. Increased use of geothermal conditioning strategies will significantly lower conventional energy consumption, peak demand, and the corresponding carbon emissions, if done with understanding of these systems and their interaction with the environment.

This entry begins with a discussion of the earth’s thermal energy and its ability to supply and store heat.

The range of geothermal conditioning approaches are then described and illustrated with pertinent examples. Available data on global use of geothermal conditioning are presented and likely future developments in the design and application of geothermal conditioning technology are discussed. Finally, printed and electronic sources of additional information about geothermal conditioning are enumerated.

The Earth’s Thermal Energy Sources and Ground Temperature Distribution

The European Union succinctly defines geothermal energy as *the energy stored in the form of heat beneath the surface of the solid earth* (RES Directive 2009/28/EC). Nothing in that definition, however, suggests the dynamic nature of the heat storage processes. Some of the short wave solar radiation incident on the surface is absorbed and conducted into soil and rock. Precipitation, either rain or snow melt that percolates into the subsurface, can transport this thermal energy substantially farther into underlying aquifers and the soil and rock surrounding them. Thermal energy is also created within the earth through the decay of radioactive isotopes such as U²³⁸ and Th²³² in granite and basalt in the earth’s crust, and through earthquake friction and the formation of new crust. These types of heat energy, in addition to thermal energy stored when the earth was formed, are transferred via conduction and convection from the interior of the earth to its surface. At the surface, long wave radiation and convective flow to the atmosphere balance the thermal energy inputs. The earth’s total heat energy content is estimated to be 12–24 10³⁰ J [5].

Within the earth’s internal energy reservoir, temperatures are estimated to reach 5,000°C (see [Table 1](#)). However, because the earth has been drilled only to a depth of 12,262 km, less than 1% of the earth’s diameter [6], and the present limit of economic drilling is approximately 4 km, the practical maximum temperature of geothermal energy is now <1,000°C. Constraining that range to temperatures effective for direct use geothermal conditioning, the range is approximately 4–150°C. An even smaller range, roughly 4–27°C, is sufficient for energy-effective building conditioning (see [Fig. 2](#)). In many areas around the world, solar radiation supplies the thermal energy for

Temperature (C)	Application
200	
190	
180	Evaporation of highly concentrated solutions, Refrigeration by ammonia absorption, Digestion in paper pulp (Kraft)
170	Heavy water via hydrogen-sulfide process
160	Drying of diatomaceous earth
150	Drying of fish meal Drying of lumber Alumina via Bayer's process
140	Drying farm products at high rates Canning of food
130	Evaporation in sugar refining. Extraction of salts by evaporation and crystallization, Fresh water by distillation
120	Most multi-effect evaporation. Concentration of saline solutions.
110	Drying and curing of light aggregate cement slabs
100	Drying of organic materials, seaweeds, grass, vegetables, etc. Washing and drying of wool
90	Drying of stock fish Intense de-icing operations
80	Space heating (buildings and greenhouses)
70	Refrigeration (low temperature limit)
60	Animal husbandry Greenhouses by combined space and hotbed heating
50	Mushroom growing Baleonology
40	Soil warming
30	Swimming pools, biodegradation, fermentations, Warm water for year-round mining in cold climates, De-icing
20	Hatching of fish. Fish farming
10	
5	
0	

Conventional power production

G

Geothermal Conditioning: Critical Sources for Sustainability. Figure 1

Original Lindal diagram: geothermal energy uses

most of that temperature range in the top 20 m of the earth's surface.

Solar Energy Input: When the sun strikes the earth's surface, the radiation must be reflected from the surface, absorbed by the surface, or transmitted through the surface to material below. For most ground surfaces, the radiation is either reflected or absorbed; some transmission occurs through snow and ice. In approximate percentages, of the total solar radiation that enters the earth's atmosphere, roughly 55% reaches

the earth's surface where 4% is immediately reflected to the atmosphere and 51% is absorbed (see Fig. 3).

The thermal impact of that absorbed radiation can be affected by dynamic processes such as wind and rain, but the static properties of the surface soil, water, and underlying rock are typically used to calculate how solar radiation affects ground temperature. The thermal *conductivity* of a material, expressed as k or λ in units of W/m/K, indicates how readily heat is transferred within a material by rapidly colliding molecules.

°C	Low Range of Conditioning Technology/Process
100	
90	
80	Cogeneration supply for district heating system
70	
60	Conventional boiler supply temperature
50	Fan coil heating water supply, Forced air furnace, Radiator water supply
40	
30	Floor heating, Ceiling panel heating, Wall panel heating
20	
10	Ceiling panel cooling water supply Sewer water supply temperature District cooling
0	Fan coil cooling water supply, Earth sheltering, Earth tube supply air Extended range heat pump entering water temperature: cooling
−10	Extended range heat pump entering water temperature: heating

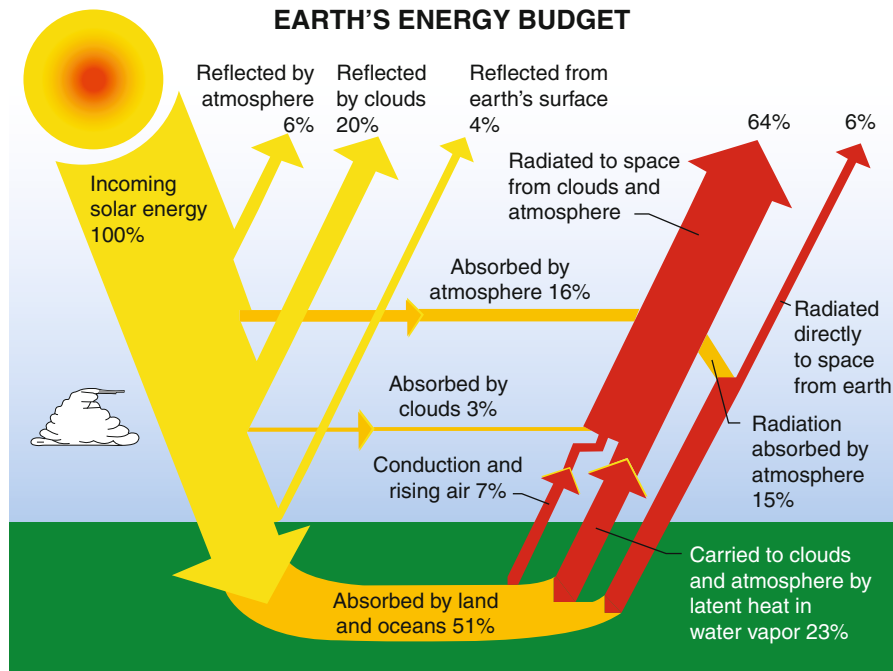
Geothermal Conditioning: Critical Sources for Sustainability. Figure 2
Low-temperature/low-exergy conditioning

Geothermal Conditioning: Critical Sources for Sustainability. Table 1 Earth temperature profile [7]

Earth layer	Depth (km)	Temperature (°C)
Near-surface solar zone	0–0.02	Similar to mean annual air temperature
Earth's crust	0–30	Up to 1,000
Earth's mantle	Up to 3,000	1,000–3,000
Earth's core	Up to 6,370	3,000–5,000

It is an important mode of heat transfer in solids although most soils and rock are relatively poor conductors. The *specific heat* of a material, expressed as c_p in units of J/kg/K, indicates how much thermal energy a material can absorb for a given change in temperature. A material's specific heat and its *density* (ρ) in kg/m³ determine its *volumetric heat capacity*, expressed

as ρc in units of MJ/m³/K. Volumetric heat capacity is the specific heat per unit volume and indicates the ability of a substance to store energy without undergoing a phase change. Soil and rock have a very high volumetric heat capacity relative to air and their ability to store heat creates a useful thermal lag. *Thermal diffusivity*, which indicates how quickly a material adjusts to the temperature of its surroundings, is calculated by dividing conductivity by volumetric heat capacity. Thermal diffusivity is expressed as κ in units of m²/s. A highly conductive material with low volumetric heat capacity will have high thermal diffusivity and readily adjust its temperature, whereas a material with similar conductivity but higher volumetric heat capacity will adjust more slowly. In contrast to high thermal diffusivity, a material with high *thermal inertia* will store heat and give it off slowly, resisting diurnal and seasonal temperature changes. Thermal inertia is the square root of the thermal conductivity, density,



Geothermal Conditioning: Critical Sources for Sustainability. Figure 3
Absorbed and reflected solar radiation

and specific heat capacity expressed as I with units of $\text{J m}^{-2} \text{K}^{-1} \text{s}^{-1/2}$. Table 2 summarizes these properties and Table 3 shows representative values for air, water, soil, and rock.

Table 3 highlights the benefits of ground- or groundwater-based conditioning compared to air-based strategies. The thermal inertia of water, soil, and rock is high. Surface temperature fluctuations diminish with depth and the time lag between the surface temperature and the ground temperature increases. Daily fluctuations are dampened within 30 cm of the surface [8]. Annually, surface temperature fluctuations reflect monthly solar cycles and attenuate with increased depth, and the time lag between the surface temperature and ground temperature is more extended with depth on an annual basis. At the depth of *zero annual range* [9] or the *neutral zone* [10], the atmosphere and soil are presumed to achieve long-term thermal equilibrium. Here, no seasonal temperature fluctuations occur and the ground temperature should reflect the long-term average annual air temperature for that particular location. Under these circumstances, the ground temperature is warmer in

winter and cooler in summer than the air temperature, which means that the ground can provide heat in winter and absorb excess heat in summer.

Table 3 also shows the impact of subsurface moisture. Saturated soil has substantially greater specific heat, volumetric heat capacity and thermal inertia than dry soil, making it a far better and more stable heat source and sink. If soil moisture varies with precipitation, the ground's thermal properties will reflect that variation. Alternatively, if soil moisture can be controlled to some extent, it may be possible to improve ground performance for geothermal conditioning.

The depth at which the neutral zone is said to occur varies within the literature (14–20 m) [9, 10] and is typically calculated, not measured. This is not surprising since determining the density, conductivity, and specific heat of heterogeneous soil and rock layers is difficult. In fact, the *ASHRAE Handbook of HVAC Applications* acknowledges that long-term field-monitored data are a major missing component in the effort to improve calculations for geothermal system design (greater detail about calculations for

Geothermal Conditioning: Critical Sources for Sustainability. Table 2 Key soil and rock properties for thermal performance

Property	Common symbol	Units	
Density	ρ	kg/m	
Specific heat capacity	c_p	J/kg K	
Thermal conductivity	k or λ	W/m K	
Volumetric heat capacity	ρc	MJ/m ³ K	Specific heat * density
Thermal diffusivity	α or κ	m ² /s	Conductivity/volumetric heat capacity
Thermal inertia	I	J m ⁻² K ⁻¹ s ^{-1/2}	$\sqrt{\text{density} * \text{specific heat} * \text{conductivity}}$

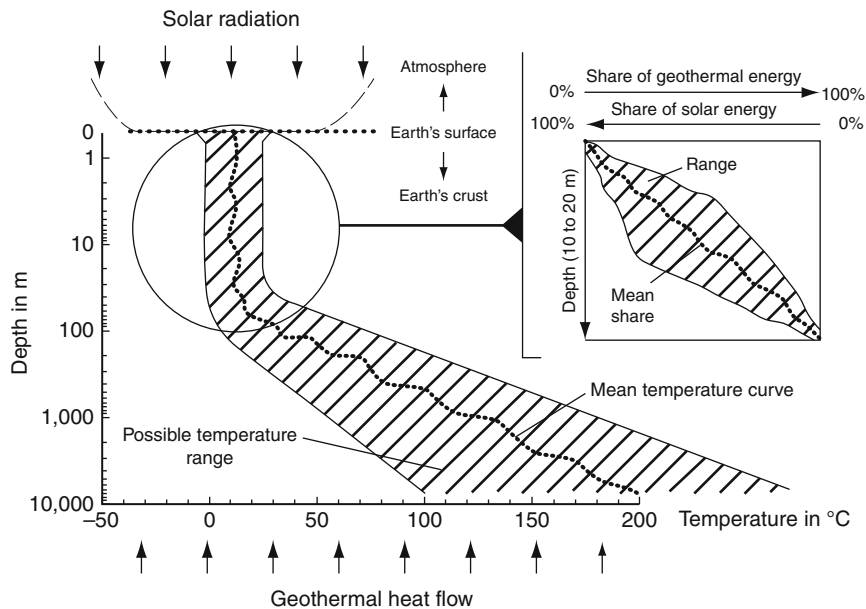
Geothermal Conditioning: Critical Sources for Sustainability. Table 3 Surface and ground properties affecting sub-surface temperature change

Material	Density (ρ) in kg/m ³	Specific heat capacity (c_p) in J/kg K	Volumetric heat capacity (ρc) in 10 ⁻⁶ J/m ³ K	Thermal conductivity (k) in W/m K	Thermal diffusivity (κ) in 10 ⁻⁶ m ² /s	Thermal inertia (I) in J m ⁻² K ⁻¹ s ^{-1/2}
Air (STP)	1.29	1,005	0.0012	0.02	15.43	5.09
Water	1,000	4,186	4.19	0.60	0.14	1,585
Sandy soil (dry)	1,600	800	1.28	0.30	0.23	620
Sandy soil (saturated)	2,000	1,480	2.96	2.20	0.74	2,552
Clay soil (dry)	1,600	890	1.42	0.26	0.18	192
Clay soil (saturated)	2,000	1,550	3.10	0.58	0.19	1,341
Rock (basalt)	2,600	800	2.08	2.50	1.20	2,280

geothermal system design is provided by Kavanaugh and Rafferty) [11]. More important than the specific depth of the neutral zone is its occurrence in the ground temperature profile. Within a few meters from the earth's surface, a zone in which building foundations exist, the ground temperature reaches or approximates long-term thermal equilibrium with the atmosphere due to incident solar radiation. Effective integration of the ground's low-temperature heat supply and heat storage capacity can substantially reduce consumption of nonrenewable energy for building conditioning.

Interior Energy Input: Beyond the neutral zone, the temperature typically rises with depth, following what is called the geothermal gradient, shown in Fig. 4 as

a broken line descending from the surface. The temperature rise along the geothermal gradient reflects heat flow from the earth's interior via conduction through solid rock and convection within geothermal fluids. Of the sources of heat energy within the earth, *radiogenic heat*, from the decay of radioactive isotopes, contributes the greatest share, annually estimated as 8.6×10^{20} J [13]. The geothermal gradient can vary substantially by locale and region. Recent research suggests that the gradient can be lowered or reversed, at least to 250 m depth, with increased energy input at the surface. Ongoing monitoring of well temperatures by Majorowicz et al. shows that rising surface air temperatures are increasing the amount of thermal energy stored in the shallow geothermal environment in



Geothermal Conditioning: Critical Sources for Sustainability. Figure 4

Ground thermal flows and temperature profile [12]

some locations. Their analysis of temperature changes down to 250 m shows that the near-surface heat gain is also creating a null or negative thermal gradient. As a result, drilling below 50 m depth in these locations offers little thermal benefit for heating applications [14]. Their findings also suggest that the ground's near-surface thermal storage (heat sink) capacity may diminish with rising air temperatures.

Globally, the average geothermal gradient ranges from 30 to 60 K/km, but the gradient may be a little as 10 K/km in older crustal areas or as high as 200 K/km in tectonically active areas such as Iceland. The most tectonically active country on earth, Iceland has more than 200 volcanoes. Nearby ground temperatures exceed 200°C at 1 km depth and outside the volcanically active zone, the geothermal gradient is still quite high at 150°C/km. In 2008, 62% of Iceland's primary energy and 24.5% of its electricity came from its geothermal resources [15].

Geothermal Conditioning Principles and Approaches

An essential principle in sustainable building design is to create a building enclosure that manages as much of the heating, cooling, and ventilation loads as possible

[see "High Performance Building Facades: Enclosures for Sustainability," V. Hartkopf, A. Aziz and V. Loftness in this volume]. With a high-performance enclosure, the size of the mechanical equipment and its related energy consumption and carbon emissions are reduced. In addition, sustainable building conditioning employs the following general strategies [16].

- Use water as an energy carrier since it carries 3,000× the energy that air carries in an equivalent volume.
- Use low-exergy systems, independently or as part of an energy cascade, to match the energy content of the supply more closely with the need.
- Use distributed rather than central systems, that is, multiple small units distributed throughout the building with commensurate controls rather than a single large central unit to cut transmission losses and to deliver thermal comfort close to building occupants.
- Zone the system to allow equipment to operate at full load efficiency and only when needed to support occupant comfort and health and building durability.
- Use renewable energy and energy recovery wherever possible.

- Design a flexible system that accommodates reconfiguration of interior space over time while maintaining original level of performance.
- Separate the delivery of ventilation air from heating and cooling, and integrate the monitoring of these systems for occupant comfort and energy effectiveness.
- Provide means for occupants to adjust thermal comfort conditions.
- Incorporate sufficient sensors and metering capability so that occupant and building operators can see, assess, and improve conditioning system performance.

Geothermal conditioning lends itself well to the application of these strategies. It provides a renewable source of thermal energy and a heat sink for heating, cooling and ventilation. With the exception of passive earth sheltering and earth tube ventilation systems, geothermal conditioning typically uses water or a water/antifreeze mix as an energy carrier for all or part of the system. Most geothermal conditioning systems are low-temperature/low-exergy systems and those that are not often use energy cascades with energy recovery to make effective use of higher temperature ground resources. The total savings from geothermal conditioning will vary widely depending on the local climate and subsurface temperatures, the conditioning strategy, and its design and operation. Nonetheless, because the ground temperature often approximates the average annual air temperature in a given location, a geothermal approach is likely to save energy wherever the average annual air temperature is closer to the building balance point temperature than the outdoor temperature range.

Discussions of geothermal conditioning often focus on heat pumps as the interior delivery system. One reason for this is that heat pumps operate efficiently at temperatures commonly accessible at and within a few meters below a building foundation. In addition, even when a heat pump's energy consumption is evaluated as source or primary energy rather than site energy, its efficiency usually exceeds 100%. (Note: Primary or source energy is the energy value based on source fuel inputs. For electricity, this is the energy value of the oil, natural gas, coal, or other source fuels used to generate the electricity and is typically 3.0–3.5 times higher than the energy value of the electricity at

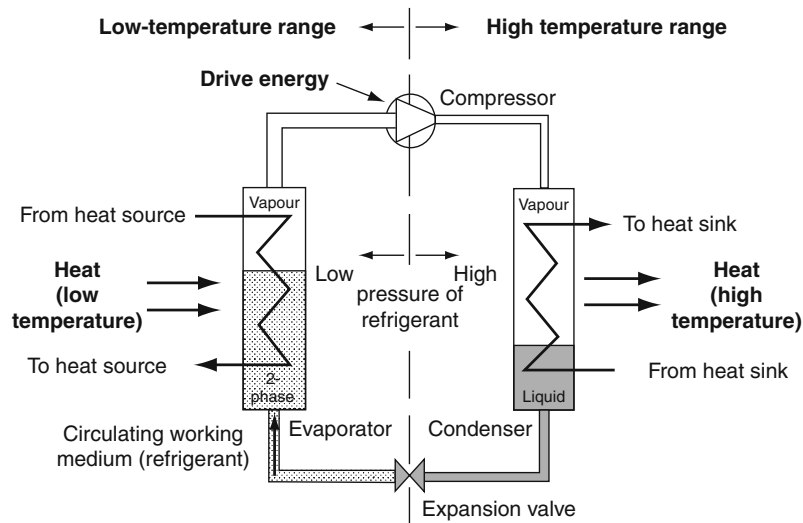
the building site.) This is because the heat pump can use heat in the ambient air, ground, or water to cause certain refrigerants to vaporize and to reach a much higher temperature when compressed. Low-grade heat coupled with a refrigerant cycle allows the heat pump to supply heat to or absorb heat from a building with a fraction of the energy required by combustion heating or conventional air conditioning.

In the case of a geothermal heat pump, the heat is being transferred from and to the ground or ground-coupled water. In heating mode, a compression heat pump transfers the earth's thermal energy, often in the range of 10–15°C, to a refrigerant that evaporates as the heat is absorbed and becomes a gas. The refrigerant gas then passes through a compressor where it is pressurized. This increases its temperature, generally above 71°C. The heated gas then passes through another heat exchanger, transferring its heat to air or water being used to heat the interior space, often at 38–43°C. As the refrigerant gas loses heat, it condenses back to a liquid, is cooled as it passes through an expansion valve, and is then ready to repeat the cycle (see Fig. 5). A reversing valve allows the heat pump to change the direction of refrigerant flow, causing heat from the indoor space to be transferred to the refrigerant and later to the ground. Compression heat pumps are by far the most common, but absorption and adsorption heat pumps also exist. Like compression heat pumps, absorption heat pumps use low-grade heat to vaporize a refrigerant, but use thermal energy rather than electricity to compress the refrigerant. Adsorption heat pumps, still in development, also use thermal energy to drive the process [17]. Heat pumps and their operation are extensively discussed in several sources [18–20], but their efficiency is a key aspect of their sustainability and is briefly summarized here.

The efficiency of the heat transfer process is the ratio of the energy used to drive the process to the amount of heat transferred. This is called the coefficient of performance or COP. In heating mode, the COP is the ratio of heat supplied to energy used. In cooling mode, the COP is the ratio of heat removed to the energy used [22].

$$\text{COP} = \frac{\text{Heat Supplied or Heat Removed (kW)}}{\text{Power Input (kW)}}$$

As an example, a geothermal heat pump that provides 12 kW of heat output with 3 kW of electricity has



Geothermal Conditioning: Critical Sources for Sustainability. Figure 5
Diagram of vapor compression heat pump in heating mode [21]

a heating COP of 4. This means that one part electrical energy and three parts ground thermal energy supply four parts of usable energy. On a site energy basis, this means that the heat pump efficiency is 400%. When considered from a source or primary energy perspective, and assuming a 30% source to site conversion efficiency for electricity, the source efficiency of this heating process is 120% ($30 \times 400\%$). The COP given by the manufacturer is determined under specific operating conditions that usually vary from those on site. Manufacturer COPs typically range from 3 to 6 at present.

The efficiency of heat pumps over the course of a year under actual operating conditions and considering auxiliary system components such as circulation pumps is expressed as the seasonal performance factor (SPF) or average annual COP. The SPF is the ratio of the system's usable energy output to the energy input.

$$\text{SPF} = \frac{\text{annual usable energy output (kWh)}}{\text{annual energy input (kWh)}}$$

This measure more accurately describes system performance and is often lower than the equipment's rated performance for several reasons. Commonly,

a heat pump that provides heating and cooling is sized for the dominant load. This may mean that a heat pump with sufficient heating capacity, for example, is oversized for cooling and therefore operates less efficiently in cooling mode. In addition, a system often does not operate at the temperatures at which it is rated but rather under more variable temperatures that decrease its efficiency. Cited SPF values for geothermal heat pumps vary considerably, but may be in the range of 3.3–4.3 [23]. Because a heat pump is able to derive a substantial portion of its heating energy or heat removal capacity from ambient air, the ground, or groundwater, its efficiency generally exceeds that of any other mechanical conditioning equipment powered by nonrenewable energy, even when the source to site conversion efficiency of the power source is much higher than that for electricity, for example, natural gas [24].

A heat pump's ability to heat and cool makes it effective in distributed systems (also known as unitary systems), particularly in mixed-use buildings. In fact, a distributed approach tends to be more energy efficient [25]. In a mixed commercial/residential building with multiple heat pumps connected to a single internal water loop, heat pumps in commercial zones with high internal loads might be in cooling mode while

those in residential zones are heating. Because a heat pump in cooling mode returns warmer water to the internal loop and a heat pump in heating mode returns cooler water to the loop, the temperature of the internal loop will be self-balancing to some extent, remaining in a temperature range favorable for efficient heat pump operation for an extended time period. This reduces the need for heat transfer between the external (ground) loop and internal loop and thus the pumping energy required. Either alone or in combination with other low-exergy equipment such as radiant wall, floor or ceiling panels or fan coil units, heat pumps can be flexibly configured to deliver thermal comfort efficiently and close to the point of use. See Fig. 6 for a geothermal heat pump system design strategy/guideline, developed by Kavanaugh [26].

Heat pumps are not, however, a component of every geothermal conditioning strategy. By focusing on the thermal properties of soil, rock, and water, a host of strategies has been developed to couple building heating and cooling with the ground and ground-coupled fluid. Tables 4 and 5 list many available options. Each capitalizes on the capacity of the ground or ground-coupled water to absorb and supply substantial thermal energy while maintaining a fairly constant temperature year-round. (Note: UTES systems are the exception; they may be designed to concentrate thermal energy for seasonal use, as described later in this section.) The basic components of each approach are (a) the ground coupling through which thermal energy is transferred to and from the ground and (b) the internal system through which thermal energy is distributed from or to the ground for comfort conditioning. Because this entry focuses on the earth's thermal energy for sustainable conditioning, Tables 4 and 5 emphasize ground coupling options. Table 4 lists closed loop systems in subsurface soil and rock and Table 5 lists water-based systems, both open and closed loops. Within the system descriptions and examples, however, internal system options are often discussed.

Like solar conditioning, geothermal conditioning can be passive or active. In essence, every building in contact with the ground exchanges heat with it, intentionally or not. In fact, any ground-based construction, including roads and pavement, alters the ground's thermal profile by changing characteristics such as albedo, reflectivity, and rainwater absorption, and by adding

heat transfer surfaces. Passive geothermal conditioning is by far the oldest approach listed, but like passive solar conditioning, it may be overlooked.

Active strategies are more numerous and can be subdivided into approaches that use subsurface soil and rock for their ground coupling and those that use water. The effectiveness of a given approach will depend on site conditions and natural resources, the local geothermal gradient, and many project-specific variables. In recent years, system designers have also developed approaches that derive their thermal stability from a subsurface location (e.g., sewer pipe) while the source of their thermal energy may be other than geothermal, for example, residual heat in building wastewater.

With the exception of earth sheltering and earth tube ventilation systems, geothermal conditioning approaches use a fluid as the energy carrier or heat transfer medium between the outdoor and indoor system components. The interface between the outdoor and indoor components is often a heat exchanger or heat pump, although other types of equipment are possible. Ground sources that provide higher temperature heat (e.g., $>30^{\circ}\text{C}$) must interface with something other than a heat pump because the maximum input water temperature for a heat pump is approximately 32°C and its efficiency is compromised at these higher temperatures. In open loop water-coupled systems, a heat exchanger is recommended as a buffer between the external water source and the indoor conditioning equipment because of the lively chemistry and/or biology of these water supplies [27].

All of the geothermal conditioning approaches in Tables 4 and 5 take advantage of the earth's stable, low-grade thermal energy and its heat capacity. Most systems use these ground characteristics for energy exchange, to provide a heat source and heat sink for the building. Some approaches target thermal storage only. Given the range of options available and an increasing emphasis on low-exergy conditioning technologies, one or more geothermal conditioning approach is probably feasible in most buildings.

Subsurface Soil and Rock Systems

With the exception of earth sheltering, subsurface soil and rock systems use a fluid enclosed in pipe to transfer

#1 - Building Layout

- * Divide floor plan into zones
- * Calculate heat loss/gain for each zone
- * Group building zones into one central or multiple ground loops

#2 - Select Equipment

- * Select heat pumps for each zone based on **capacity and efficiency** at design conditions
- * Consider head loss, temperature range, package type, sound, serviceability
- * Specify water source water heating and refrigeration equipment if applicable
- * Select ventilation air system components—ducting, heat recovery, preconditioning coils, etc.

#3a - For GCHPs

- * Determine ground properties (test bores)
- * Specify tube type, size, bore separation, backfill
- * Calculate required bore
- * Design exterior headers
- * Design purge system

3b - For GWHPs

- * Determine groundwater availability/quality
- * Specify required well flow
- * Specify water disposal method
- * Specify groundwater-to-loop water heat exchanger

3c - For SWHPs

- * Find reservoir flows, depth, and temperatures (high/low)
- * Specify coil size & type
- * Calculate required coil length
- * Design exterior headers
- * Design purge system

#4 - Design Building Piping Loop

- * Weigh advantages of central loop vs. multiple loops
- * Route and size piping system for low pressure losses
- * Provide on-off flow control through heat pumps and isolation valves
- * Specify materials—indoor piping, insulation, antifreeze, inhibitors

#5 - Specify Pump and Control Method

- * Weigh advantages of central pump(s) vs. multiple remote pumps
- * Select pump(s) to operate near maximum efficiency on pump curve
- * Weigh pump control options—no control, on-off, multispeed (or multiple pump), variable speed
- * Calculate loop pump power and redesign system if greater than 10% of total demand

#6 - Evaluate Other Alternatives

- * Use higher efficiency heat pumps to reduce required ground loop size?
- * Use cooling tower or fluid cooler to reduce loop size?
- * Increase or decrease bore separation or coil tubing size?
- * Look at cost of multiple loops and pumps vs. central loops and pump—**include cost of controls.**

Geothermal Conditioning: Critical Sources for Sustainability. Figure 6

Strategy for geothermal heat pump system design. *GCHPs* ground-coupled heat pumps (soil-based systems), *GWHPs* groundwater heat pumps, *SWHPs* surface water heat pumps

Geothermal Conditioning: Critical Sources for Sustainability. Table 4 Subsurface soil and rock systems (all closed loops)

Passive or active strategy	Exterior equipment/configuration	Typical depth	Energy carrier	Function	Exterior/interior interface
Passive	Earth berms	At grade or foundation depth	Soil	Insulation, thermal mass, reduced infiltration, reduced solar gain	Building enclosure
Active	Earth tubes	3–5 m	Air	Preheating or cooling of ventilation air	Energy recovery ventilator; possibly UV disinfection
	Horizontal trench with straight pipe or slinky coil	1–2 m	Water–antifreeze mix	Energy exchange for heating and/or cooling	Heat pump
	Direct connection to conditioning equipment	1.2–2 m	Refrigerant		Heat pump
	Building foundation piles (open or closed)	5–30 m	Water or water–antifreeze mix		Heat pump or heat exchanger
	Shallow vertical boreholes with U-tube pipe	30–120 m	Water or water–antifreeze mix		Heat pump
	Deep vertical boreholes with coaxial pipe	1,000–3,000 m	Water		Heating equipment or heat exchanger (typically cooling not attempted with deep systems)
	BTES (borehole thermal energy storage)	20–300 m	Water or water–antifreeze mix	Seasonal hot or cold storage for increased efficiency	Heat pump, heat exchanger, cogeneration power plant, solar hot water system

heat between the building and the outdoors. For earth tubes, that fluid is air and for other systems, refrigerant, water, or a water/antifreeze mix. Under normal operating conditions, there is no direct contact between the fluid and the subsurface environment and this type of sealed subsurface pipe heat exchanger is called a closed loop system.

Design and sizing of the ground heat exchanger is a key aspect of system cost and performance. The characteristics of underlying soils and rock, ground moisture content and water movement are among the variables that affect ground temperature and heat

transfer and these can be difficult to model with precision. Research has also shown that in low-temperature systems, the effectiveness of the ground heat exchanger is strongly associated with its heat transfer over time, particularly when multiple heat exchanger pipes are in close proximity [28]. Abundant research focuses on this topic and many sizing calculations, modeling tools, and rules of thumb exist to support this task [13, 29]. It is worth noting, however, that the validity of available sizing methods is still debated by those who design and operate closed loop systems [30]. Data from operating systems are also lacking [31]. Despite these

Geothermal Conditioning: Critical Sources for Sustainability. Table 5 Water-based systems (all active strategies, both closed and open loops)

Water source	Open or closed loop	Exterior equipment/ configuration	Typical depth	Energy carrier	Function	Exterior/ interior interface
Aquifer	Open	Supply well with surface discharge	15–100	Groundwater	Energy exchange for heating and/or cooling	Heat exchanger
		Supply and injection wells, shallow or deep	15+	Groundwater		Heat exchanger, cooling or heating equipment
		Standing column (or coaxial) well	120–375 m	Groundwater		Heat pump
		ATES (aquifer thermal energy storage)	10–400 m	Groundwater	Seasonal hot or cold storage for increased efficiency	Heat exchanger
Surface water (ocean, lake, river)	Open	Supply and reinjection pipe	Varies by water body; ideally at depth where temperature is stable and where reinjection does not affect natural seasonal cycling	Surface water	Energy exchange for heating and/or cooling	Heat exchanger
	Closed or Open	Slinky coil	3–3.5 m minimum; sufficient depth so that supply temperature always $>4^{\circ}\text{C}$	Water or water–antifreeze mix		Heat pump or heat exchanger
Abandoned mine tunnels	Open	Supply and reinjection pipe	Varies with mine	Groundwater	Energy exchange for heating and/or cooling	Heat exchanger, heating or cooling equipment
	Open	CTES (cavern thermal energy storage)	Up to 2,000 m	Groundwater possibly with phase change material	Seasonal hot or cold storage for increased efficiency	
Sewer infrastructure	Closed or Open	Heat exchanger within/around pipe	Standard sewer pipe depth for given location	Water or water–antifreeze mix	Energy exchange for heating and/or cooling	Heat pump
Water treatment infrastructure	Open	Supply and reinjection pipe	Surface or subsurface pipe	Treated wastewater	Energy exchange for heating and/or cooling	Heat exchanger

challenges, the majority of ground-coupled systems now being installed are closed loop systems [32]. Brief descriptions of soil- and rock-based systems and installed examples are offered below.

Earth Sheltering The simplest form of geothermal conditioning is the passive strategy of earth sheltering, also called earth berming or earth integration. Ground temperatures begin to converge on a steady value just 30–50 cm below the surface [8]. An earth-sheltered building enclosure at and below this depth experiences temperatures that vary only a few degrees throughout the year and will generally be warmer in winter and cooler in summer than the outdoor temperature. The constant temperature, together with the reductions in air leakage and in convective and conductive heat loss that accompany earth sheltering, can substantially reduce mechanical conditioning requirements. The effectiveness of earth sheltering is well established, but its use requires careful attention to orientation, structure, waterproofing, daylighting, and egress as well as the standard range of design considerations. Although current enclosure design strategies focus largely on reducing heat transfer at the foundation, there may be substantial room for development or redevelopment of passive geothermal conditioning strategies [33].

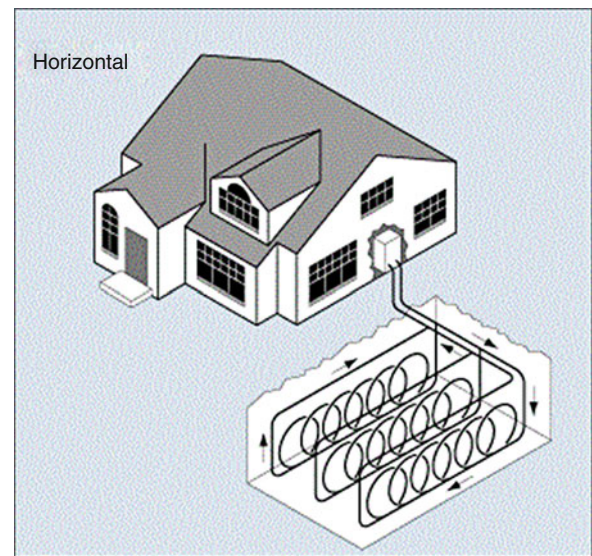
Earth Tubes (also called Earth Pipes or Earth-to-Air Heat Exchangers)

Earth tubes are buried pipes through which outdoor air is drawn into the building. The earth's temperature can heat the air in winter and cool it in summer, reducing the energy required to condition ventilation air and in some instances, eliminating the need for additional cooling equipment. Monitoring data for earth tube systems are still limited, but Pfafferoth [34] offers a comparative analysis of the performance of systems installed in three buildings in Germany between 1999 and 2001. He found that each system supplied far more heating and cooling energy than the primary energy used by the fans and that the ground characteristics and the impact of the building on ground temperature were as important as the earth tube diameter on thermal efficiency. To avoid unwanted heating in summer and cooling in winter, a control strategy is necessary. In some systems,

permeable pipe is used to allow condensation within the tubes to evaporate and UV filtration is used to address mold or bacteria that may be in the air stream. Where radon is a concern, care must also be taken that the earth tubes are not transferring radon gas from the ground into the ventilation system [35].

Horizontal Trench, Pipe, or Slinky Coil Small building loads can be handled with a horizontal ground heat exchanger installed below the frost line (Fig. 7). Since trenching is less expensive than drilling, horizontal installations cost less but require more land area; increased temperature fluctuations at shallower depths (1–2 m) mean that increased pipe length (heat transfer surface area) is required. Because of the land area requirement, horizontal ground heat exchangers typically serve building loads less than 175 kW.

Building Foundation Heat Exchangers (foundation piles, energy piles, slot-die walls): Structural components such as piles, retaining walls, and foundation slabs can be used to exchange thermal energy between



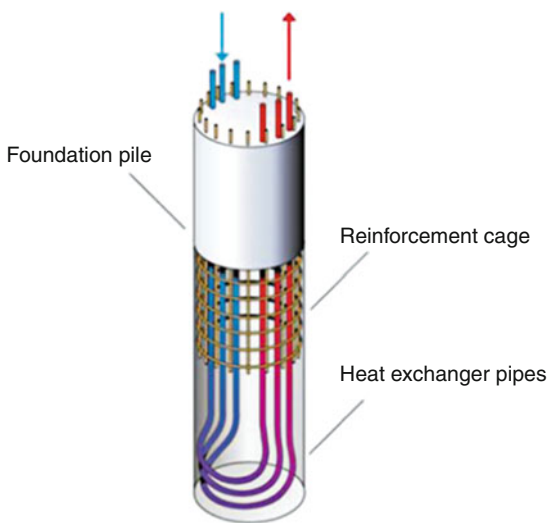
Geothermal Conditioning: Critical Sources for Sustainability. Figure 7

Horizontal ground heat exchanger

the ground and the building. Fluid-filled pipe systems are incorporated inside the foundation elements and serve a purpose similar to vertical borehole exchangers, but within the building's footprint. Relatively new (1990s) as a design strategy, foundation piles may be steel, precast concrete, or cast in place concrete [36]. The piles contain two or more U-tubes (see Fig. 8) that are connected either directly (open) or indirectly (closed) to the building's mechanical system. Pump energy is minimized since the heat exchanger is within the building footprint. Currently, design of these systems is adapted from vertical borehole design and because of uncertainties about the mechanical behavior of the soil over time (thermoelasticity and soil strength with heating and cooling), substantial safety factors are built into these systems, resulting in high costs [37].

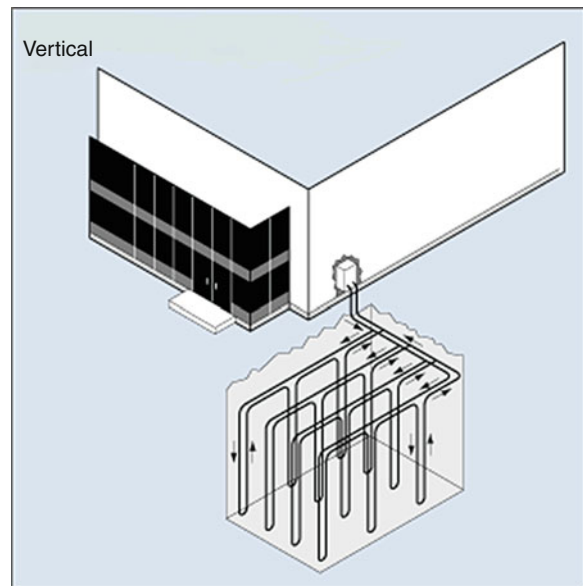
Vertical Boreholes, Shallow The most common type of closed loop ground heat exchanger is a shallow vertical borehole system. Borehole diameters vary by pipe size and by construction convention, which differs by country to some extent. In the United States, a typical borehole is about 10–15 cm diameter. The depth may range from 30 to 120 m, depending on several factors such as the conditioning load, available land area, ground conditions and temperature, and drilling

costs. A U-tube or ground probe, usually of high-density polyethylene (HDPE) pipe and often 19–25 mm pipe diameter, is inserted into each borehole (see Figs. 9 and 10) and the borehole should be packed and sealed with grout, at least at the surface, to prevent surface contaminants from passing easily down the borehole and into groundwater. (Note: Requirements for grout vary, although the potential for surface contaminants to reach groundwater more easily through a borehole does not. Grouts with improved heat transfer characteristics are available.) The total underground pipe length must provide sufficient heat transfer to satisfy the connected peak block load, which is the maximum cooling or heating load, whichever is greater, imposed on the conditioning equipment during the cooling or heating season. (Note: If the building has good load diversity, the peak block load will be less than the sum of the peak room loads or peak zone loads.) Except where borehole thermal energy storage (BTES) is used, boreholes are often spaced 4.5–6 m apart so that heat transfer with the ground is not compromised by proximity to other boreholes. A radial configuration for shallow boreholes has also been developed [39].



Geothermal Conditioning: Critical Sources for Sustainability. Figure 8

Building foundation pile heat exchanger [38]



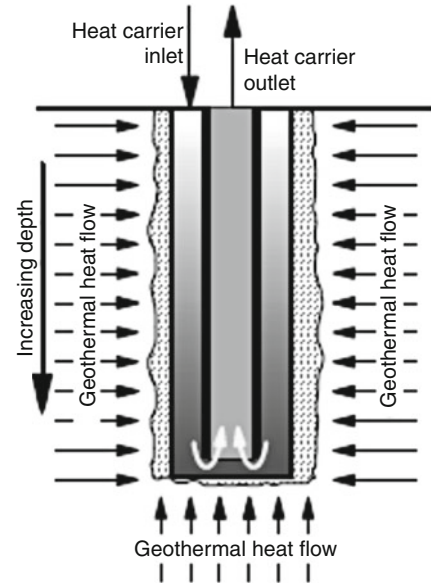
Geothermal Conditioning: Critical Sources for Sustainability. Figure 9

Vertical borehole heat exchanger



Geothermal Conditioning: Critical Sources for Sustainability. Figure 10

HDPE U-tube for vertical borehole [40]



Geothermal Conditioning: Critical Sources for Sustainability. Figure 11

Deep borehole heat exchanger [42]

Vertical Boreholes, Deep Deep borehole systems, 1,000–3,000 m, capitalize on higher ground temperature at depth and so are predominantly used for heating. In this type of system, an insulated production pipe is inserted into a borehole casing sealed at the bottom, providing a concentric pipe configuration. The heat transfer fluid, typically water that may be treated with a corrosion inhibitor, is pumped down the borehole casing, and up through the central production pipe (see Fig. 11). There is no direct contact between the heat transfer fluid and the subsurface soil and rock; the fluid gains heat at depth according to the geothermal gradient. At well exit, the heat transfer fluid passes through a heat exchanger, heat pump, or other heating equipment, depending on the fluid temperature ($\pm 40^\circ\text{C}$). The higher cost of deep borehole systems makes them better suited to serve the base load in systems that have sizable heat demand (see Fig. 12) [41].

Borehole Thermal Energy Storage Borehole thermal energy storage, BTES, is a type of underground thermal energy storage (UTES) system used to concentrate heat and cold for seasonal use. For a BTES, vertical

boreholes are placed in close proximity (1.5–3 m apart) to seasonally charge the system with hot or cold fluid. In summer, for example, a heat pump or other mechanical device can transfer heat to the ground for winter heating. In winter, that heat can be extracted and the ground can be charged with cold fluid for summer cooling. For efficiency, different ground areas or depths are often used for hot and cold storage. Although some heat or cold is lost at the system edges and at the ground surface, these systems can be used successfully in existing subsurface rock and soil; constructed storage vessels are not required. From a cost standpoint, underground thermal energy storage systems, including BTES, ATES, and CTES systems, tend to work best for small district systems rather than for single buildings.

Examples of Subsurface Soil and Rock Geothermal Systems *Earth Tubes:* Fraunhofer ISE, Freiburg, Germany: Installed in 2001, the earth-to-air energy exchanger provides cooled or preheated air with a ground temperature of 13.8°C . This open loop system consists of seven polyethylene ducts, 90–100 m long and 250 mm in diameter, buried 4–5 m and



District system served by deep borehole exchanger [41]

combined into 16 in. (40.64 cm) supply and return lines that serve heat pumps (35–123 kW) in campus buildings. The system reduced natural gas consumption approximately 75% and electricity consumption 25%, despite the switch to heat pumps. Since 1993, the college has added substantial geothermal conditioning infrastructure, both closed and open loop systems, and in 2008, an aquifer thermal storage system [43]. Extensive monitoring of system performance and ground/groundwater conditions is conducted.

Building Foundation Piles: The nursing school at Sapporo City University in Japan, completed in 2006, uses steel foundation piles filled with water to provide a ground heat exchanger that supplies the building's base heating and outdoor air cooling loads. For a floor area of 2,800 m², 51 piles from 600 to 800 mm diameter were drilled 4 m into the ground. Two U-tubes were inserted in each pile and the pile was filled with 115 m³ water. The manifold system for the piles feeds a 50 kW heat pump that provides both heating and cooling. Three vertical boreholes 75 m long supplement the system so that the ground heat

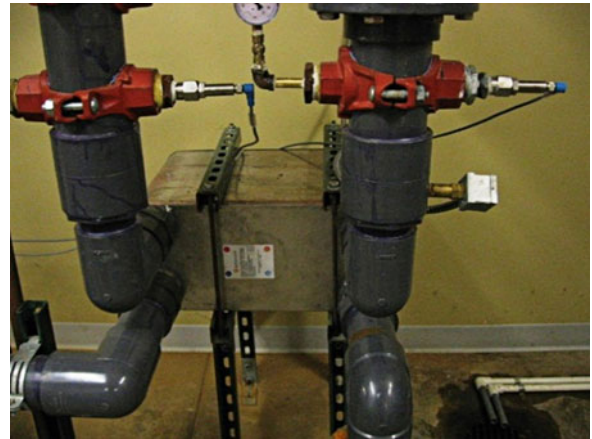
exchanger system can meet the base heating load of 50 kW and the base cooling load for outside air. During 2007, the piles extracted 43,929 kWh of heat from the ground in winter and transfer 53,939 kWh of heat into the ground in summer [36].

Borehole Thermal Energy Storage (BTES): The Drake Landing Solar Community in Okotoks, Alberta (Canada) relies on the earth's high volumetric heat capacity to store heat captured with solar hot water systems in summer months and to provide heat throughout Alberta's cold winter. Opened in 2007, the Drake Landing project is a district system serving 52 well-insulated single-family homes. Heat from 800 single-glazed roof-mounted solar hot water panels is fed to a seasonal thermal storage system that consists of 144 vertical boreholes approximately 30 m deep. The borehole field was anticipated to require 5 years to fully charge to 80°C by summer's end, thereby allowing it to supply 90% of the community's heating energy (90% solar fraction). However, 3 years after it began operating, the system reached 80% solar fraction and should be fully charged in 2011. Fan coil units are the terminal heating equipment in each home [44].

Water-Based Systems

Water-based systems use water that originates outside the building – generally groundwater or surface water – as the energy carrier. With a few exceptions, this externally sourced water is piped to the building where it transfers heat in direct contact with some component of the conditioning system: a heat exchanger, heat pump, or other conditioning equipment (see Fig. 13). It is then piped away from the building for release to or near the original water source and usually at a temperature within 6°C of its source temperature. This is called an open loop system. Closed loop water-based systems bring building piping to the external water source such as a pond or lake. Like other closed loop systems, the building piping is filled with a fluid that transfers heat between the building and the outdoors and there is no direct contact between the external water source and the heat transfer medium in the building pipes.

The reason for emphasizing direct contact between an external water supply and the building mechanical



Geothermal Conditioning: Critical Sources for Sustainability. Figure 13

Brazed plate and frame heat exchanger in 4,000 m² mixed-use building. Water's remarkable heat capacity allows smaller mechanical system components than those required by air-based systems [47]

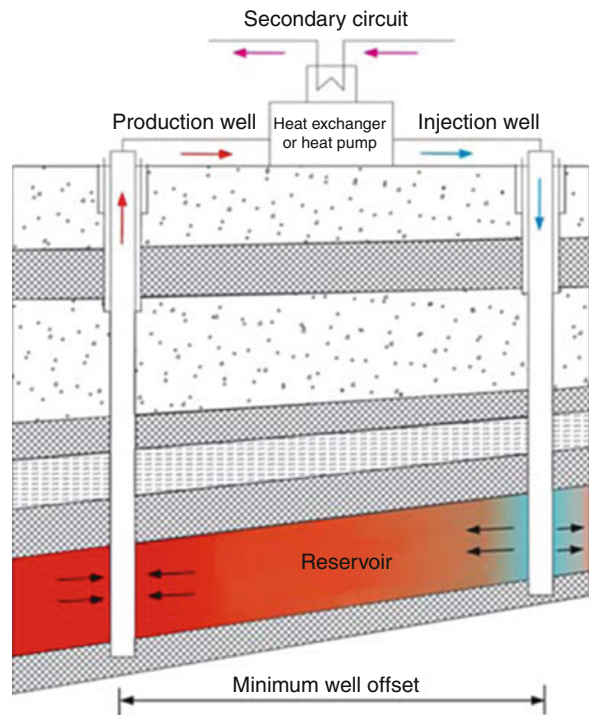
equipment, or the lack thereof, is that water is an excellent solvent that “dissolves more substances in greater quantities than any other liquid” [45]. Water will react with pipes and equipment, and natural water such as groundwater or surface water contains minerals and microbes from surrounding soil and rock that can promote those reactions, particularly in the presence of oxygen (see Fig. 15). Whereas equipment connected to the municipal water supply is often treated with chemicals to control chemical and biological activity (e.g., cooling towers), equipment connected to a natural water supply cannot be handled the same way. This characteristic is understood and addressed in water well literature, but underemphasized in building systems literature and should not be ignored in practice when sustainability is the goal [46].

As with subsurface soil and rock systems, the design and sizing of the heat exchange system is a key aspect of overall cost and performance. One obvious prerequisite is the proximity of a suitable water source. Local regulations must also support the use of such systems. In groundwater applications, the water is pumped from an aquifer at stable temperatures similar to the ground temperature at that depth and is discharged either to

surface water or to the same aquifer. Ease of access, water quantity, water quality, and viable discharge options are the limiting factors. In surface water applications, which can be either open or closed loops, water temperature and quantity are essential factors for effective heat exchange. The water temperature must not fall below 4°C and must remain in an efficient range for cooling (preferably <23°C) regardless of solar radiation and other heat transfer to the water. For open loop surface water applications, water quality is also important and can vary seasonally and during storms. Whereas soil- and rock-based heat exchangers are estimated to have a heat output of 20–50 W/m², the heat output of groundwater and surface water systems is estimated to be 2,300–4,600 W/m³/h [18]. Brief descriptions of water-based systems and installed examples are offered below.

Aquifer-Based System Before the development of plastic pipe that could be used for closed loop systems, aquifer-based open loop geothermal systems predominated. Two primary configurations exist: those that reinject the water into the ground, usually into the source aquifer (Fig. 14), and those that discharge to surface water (pump and release or, colloquially, pump and dump systems). Well depths may range from 15 to 100+ m and flow from 1 to 125+ L/s.

Because they require less exterior infrastructure, aquifer-based systems need less land area and have lower first costs. Conventionally, the considerations in choosing an aquifer-based system include proximity to the building site; groundwater availability/quantity, depth and flow rate; water chemistry; the conditioning system temperature drop and load factor; and the groundwater discharge strategy. Today, however, a primary consideration for the sustainability of these and all other groundwater applications may be the protection of drinking water resources. Groundwater is a preferred source of drinking water, but lax or nonexistent groundwater regulation in the past and even at present has allowed contamination and depletion of known aquifers. Population growth compounds these problems. Although some aquifer-based systems have been operating for decades without dropping the groundwater table, ongoing use of aquifer water for any application must consider its impact on future drinking water supplies.



Geothermal Conditioning: Critical Sources for Sustainability. Figure 14

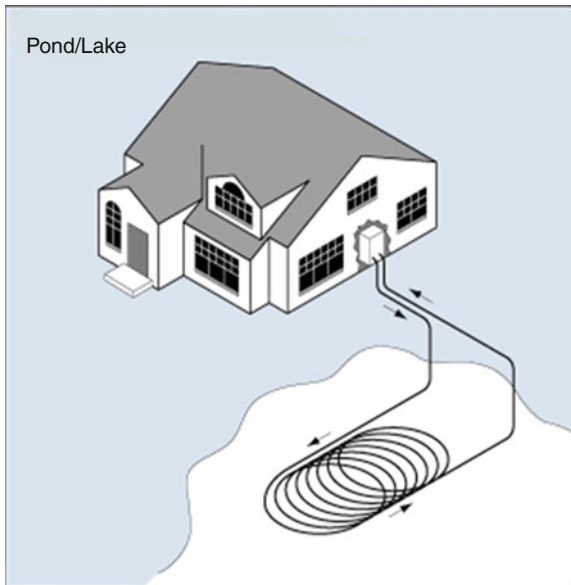
Open loop aquifer system with production and injection well [48]

Standing Column or Coaxial Well A standing column well (SCW) is an open loop system that draws and reinjects to the same well. SCWs are generally used where bedrock is close to the surface. The ground heat exchanger is a vertical well typically 150 mm diameter in bedrock with a 200 mm steel casing in the overlying unconsolidated layer. In shallower wells, the submersible pump is at the bottom of the well with the return near the top. In wells deeper than 150 m, return is generally via a 100 mm PVC dip tube that extends to the well bottom and is perforated for the last 6–18 m. A distinct characteristic of these wells is the practice of “bleeding” the well to control the temperature. Where freezing is possible, the well may be bled, typically with surface disposal, to draw warmer water into the well. Bleeding may also be used if the well is overheating in summer. SCWs connect directly to a geothermal heat pump and in residential systems may also provide domestic water.



Geothermal Conditioning: Critical Sources for Sustainability. Figure 15

Heat exchanger fouling from excessive iron in groundwater [49]. Failure to test groundwater prior to system installation may result in a system too maintenance-intensive to operate economically



Geothermal Conditioning: Critical Sources for Sustainability. Figure 16

Closed loop surface water exchanger

Surface Water System The diverse thermal profile of surface water bodies results in similarly diverse strategies for surface water geothermal conditioning. Surface

water systems can be either open or closed loops. They can be as shallow as a pond and as deep as the ocean (e.g., the Hawaii Gateway Energy Center, described below). In closed loop systems, water or a water/anti-freeze mix is pumped through a submerged pipe loop that transfers heat to and from the body of water (see Fig.16). Compared to soil-based closed loop systems, water-based systems have lower excavation costs and usually lower costs for pumping and for overall operation and maintenance. Temperature variations in surface water are typically far greater than those in soil, however, so the system may be less efficient and at greater risk for freezing and other damage if the piping is accessible and/or close to the surface. In open loop systems, the surface water body is often used for a heat sink, similar in function to a cooling tower but without the noise, fan energy, chemical dosing, and maintenance demands. Direct cooling or precooling of ventilation air is also possible by pumping water at 10°C or below through a coil within a return air duct or convector system.

Aquifer Thermal Energy Storage Similar to BTES, aquifer thermal energy storage (ATES) is a means to concentrate heat and cold underground for seasonal use. A series of injection wells is used to inject either heated (13–120°C) or cooled (6–12°C) water for later withdrawal. Heat and cold can be stored at different depths if the underlying geology permits. Withdrawal can occur through the same wells, or through separate ones, in which case the water flows in the aquifer between injection and supply. China began installing large open ATES systems (surface water allowed to infiltrate groundwater) for cold storage in the 1960s [50]. The longest operating high-temperature system was installed at Utrecht University in the Netherlands in 1991, using residual heat from cogeneration. Estimates of energy savings from ATES are quite high: 80% reduction in cooling costs and 40% reduction in heating costs [51]. Research indicates that a large percentage of land area is underlain by aquifers that could be used for ATES [51].

Abandoned Mine Tunnel/Cavity Systems Abandoned underground mines and tunnels accumulate groundwater. Depending on their size, these cavities can thermally function like an aquifer, providing an

underground reservoir at a stable temperature. Cavity thermal energy storage (CTES) systems typically use supply and injection wells. In some cases, environmental regulation (e.g., in the USA) require that groundwater in mine tunnels be treated for surface discharge. Where this occurs, a water treatment infrastructure system (see below) may be possible.

Sewer and Water Treatment Infrastructure Systems

Sewer pipes and municipal or industrial pipes that carry treated water to its discharge point are typically buried and benefit from the ground's thermal stability and capacity. In addition, the processes that serve these pipes add heat energy so that the fluids they carry often have thermal energy content comparable to or higher than groundwater ($>10^{\circ}\text{C}$). Although water treatment pipes may not be colocated with the buildings they serve, sewer pipes are. With constant and sufficient flow, these pipes and/or the fluids they carry provide a heat source and heat sink for geothermal conditioning without the need for new ground infrastructure. At least one company manufactures concrete sewer pipe with embedded heat exchangers to permit noncontact heat transfer (see Fig. 17) [52].

Examples of Water-Based Geothermal Systems

Aquifer Geothermal System with Surface Discharge: The Galt House Hotel and Waterfront Plaza Office

Towers in Louisville, Kentucky (USA) use a 21 MW open loop heat pump system to condition approximately $185,806\text{ m}^2$ of hotel, office, and apartment space (see Figs. 18 and 19). The system was initially installed in 1984 and has since expanded to its current capacity. Seven wells approximately 40 m deep supply groundwater at 14°C and approximately 46 L/s to a central network of seven plate and frame heat exchangers. The groundwater discharges to the Ohio River. Distributed heat pump loops serve the buildings' interior. Because the well pumps are close to the ground surface, the groundwater system is shut down when outdoor temperatures fall below 4°C and boilers supply the building loop during cold weather.

Seawater Direct Cooling: The Hawaii Gateway Energy Center (HGECC) visitor center on the south coast of Kona on the Big Island of Hawaii pumps seawater at 7°C from 914 m below the surface to cool the 334 m^2 building. The water is distributed through cooling coils in a subfloor plenum, absorbing heat from air supplied to the plenum through a dedicated air inlet structure. The conditioned air then rises through the building to thermal chimney outlet pipes for exhaust. This open loop seawater cooling system runs continuously throughout the day, providing 10–15 air changes per hour. Cooling



Geothermal Conditioning: Critical Sources for Sustainability. Figure 17

Sewer pipe with embedded heat exchanger [52]



Geothermal Conditioning: Critical Sources for Sustainability. Figure 18

Original Galt House Hotel on left; Hotel expansion on right, Louisville, KY (USA) [40]



Geothermal Conditioning: Critical Sources for Sustainability. Figure 19

Waterfront Plaza Office Towers, Louisville, KY [40]

coil condensate is collected and used for irrigation and toilet flushing. No auxiliary cooling equipment is used; only pump energy is consumed. HGEC produces more electricity than it uses (a net energy exporter) and the pump energy is supplied by a PV array.

Aquifer Thermal Energy Storage (ATES): At the Reichstag Building in Berlin, Germany, two confined aquifers at different depths are used for thermal separation and storage. The shallower aquifer ranges from 0 to 66 m depth and is served by two sets of five wells. The deeper aquifer lies at 270–370 m depth and is served by two wells 300 m apart. In summer months, residual heat from two combined heat and power (CHP) plants in the building charges the lower aquifer to approximately 70°C. During the winter, this water directly serves heating equipment at the beginning of winter and later, as the storage temperature drops during the heating season to 45°C, its temperature is boosted with an absorption heat pump. Meanwhile, the upper aquifer is charged with ambient cold during winter to a temperature of

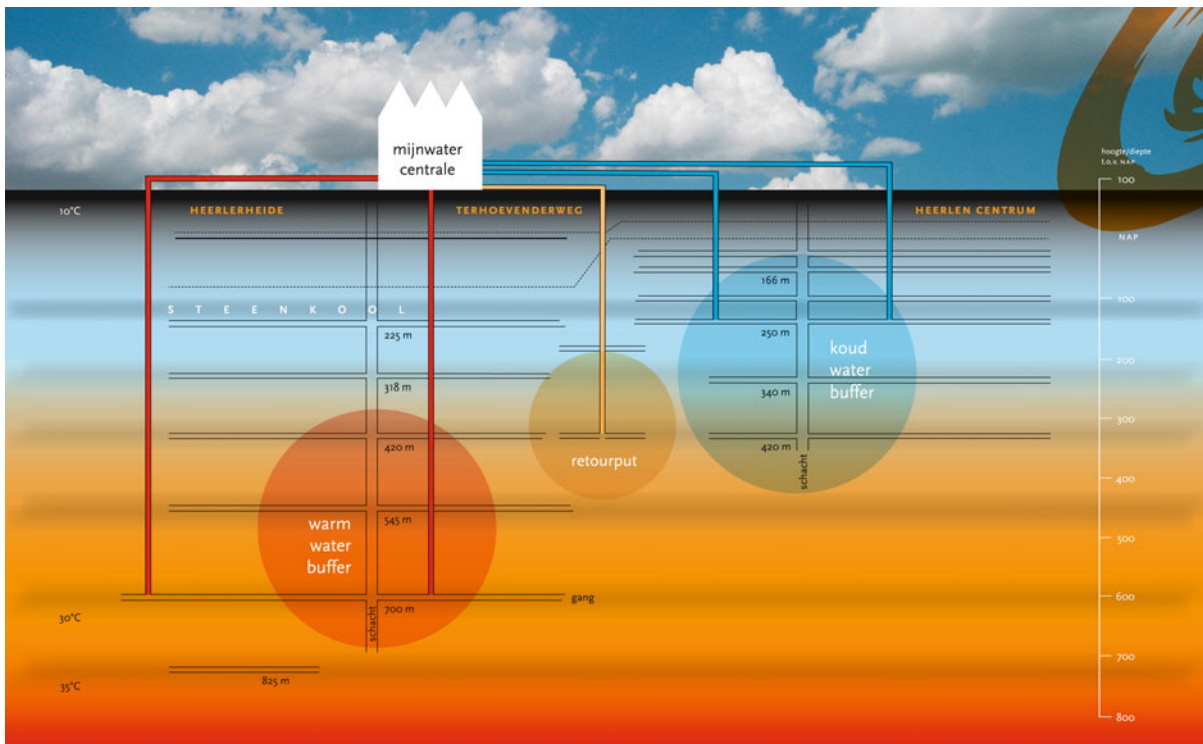


Geothermal Conditioning: Critical Sources for Sustainability. Figure 20

Cultural Center: “Gen Coel” in Heerlen – home of mine tunnel geothermal district system [55]

10°C. In the summer, cooling needs are met with the cold storage and with an absorption heat pump serving as a chiller [53].

Mine Tunnel Geothermal District System with Cavern Thermal Energy Storage (CTES): In Heerlen, the Netherlands, flooded coal mine tunnels abandoned about 1960 are now part of a district geothermal system that provides conditioning to 350 homes and to businesses (see Figs. 20 and 21). Five wells 700 m deep serve a primary energy grid. Each well can supply almost 80 m³/h at approximately 32°C. At distributed local “energy stations,” heat exchangers transfer energy to the secondary grid that serves district buildings. Heat pumps, combined heat and power (CHP) equipment, and condensing gas boilers are used to add thermal energy for conditioning and domestic hot water, depending on the buildings’ requirements. The system supplies low-temperature heating (35–40°C), high-temperature



Geothermal Conditioning: Critical Sources for Sustainability. Figure 21

Schematic diagram of Heerlen subsurface system [55]

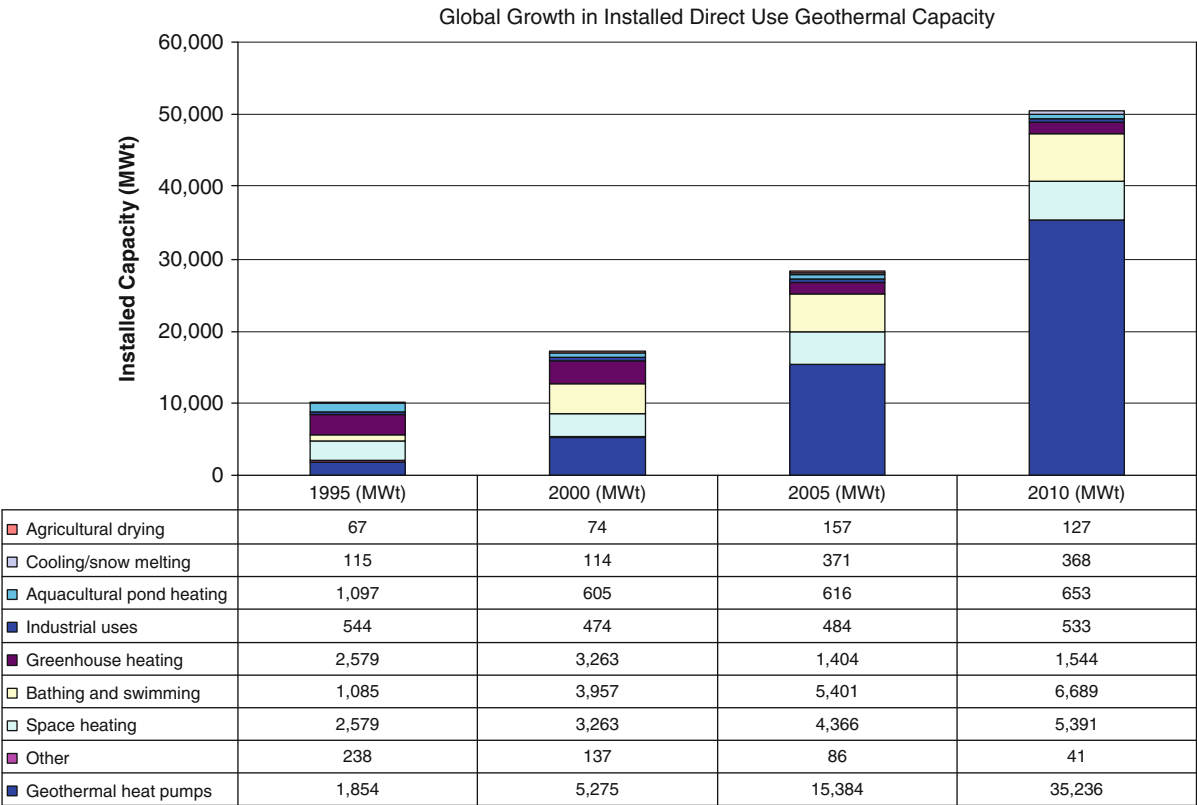
cooling (16–18°C) supply, and a combined return (20–23°C). The system includes both warm and cool thermal energy storage at 450 and 250 m depth, respectively [54].

Water Treatment Infrastructure: To supply a new geothermal heat pump district system, Oceana Naval Air Station in Dam Neck Annex, Virginia (USA) taps a pipe that carries 113,592 m³ of 21°C treated wastewater to the ocean daily. The Hampton Roads Sanitary District (HRSD) owns the 1.6 m diameter concrete reinforced pipe that runs above and below ground across Navy property carrying the treated water 2.4 km into the Atlantic Ocean. In a planned conversion from a central steam plant to a 1.6 MW geothermal heat pump coupled with a 14.5 MW cooling water condenser loop, the Navy avoided the installation of a 2,100 vertical borehole ground heat exchanger by using the treated wastewater. Almost 53,000 m³ of the wastewater flow through four plate and frame heat exchangers daily

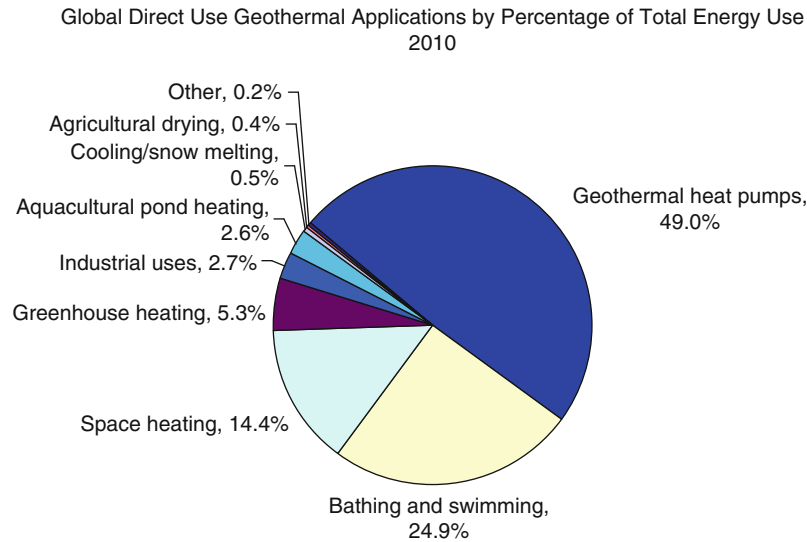
and are returned to the pipe within 1.7°C of the supply temperature. Sixty-five percent of the base is currently on the new system, which is expandable. In its first year of operation (2009), the system reduced energy consumption for building conditioning by 40%.

Installed Capacity and Annual Energy Use

Global geothermal conditioning capacity including earth sheltering, earth tubes, and ground-based heat exchangers connected to heat pumps or other indoor conditioning equipment is not fully tracked. However, the World Geothermal Congress uses national reports to estimate the total thermal power of direct use geothermal systems. In an analysis of those reports, Lund, Freeston, and Boyd state that installed capacity increased almost sixfold between 1995 and 2010, from 8,664 to 50,583 MWt [56]. Most recent data show that approximately 49% is for



Geothermal Conditioning: Critical Sources for Sustainability. Figure 22
Global growth in installed direct use geothermal capacity



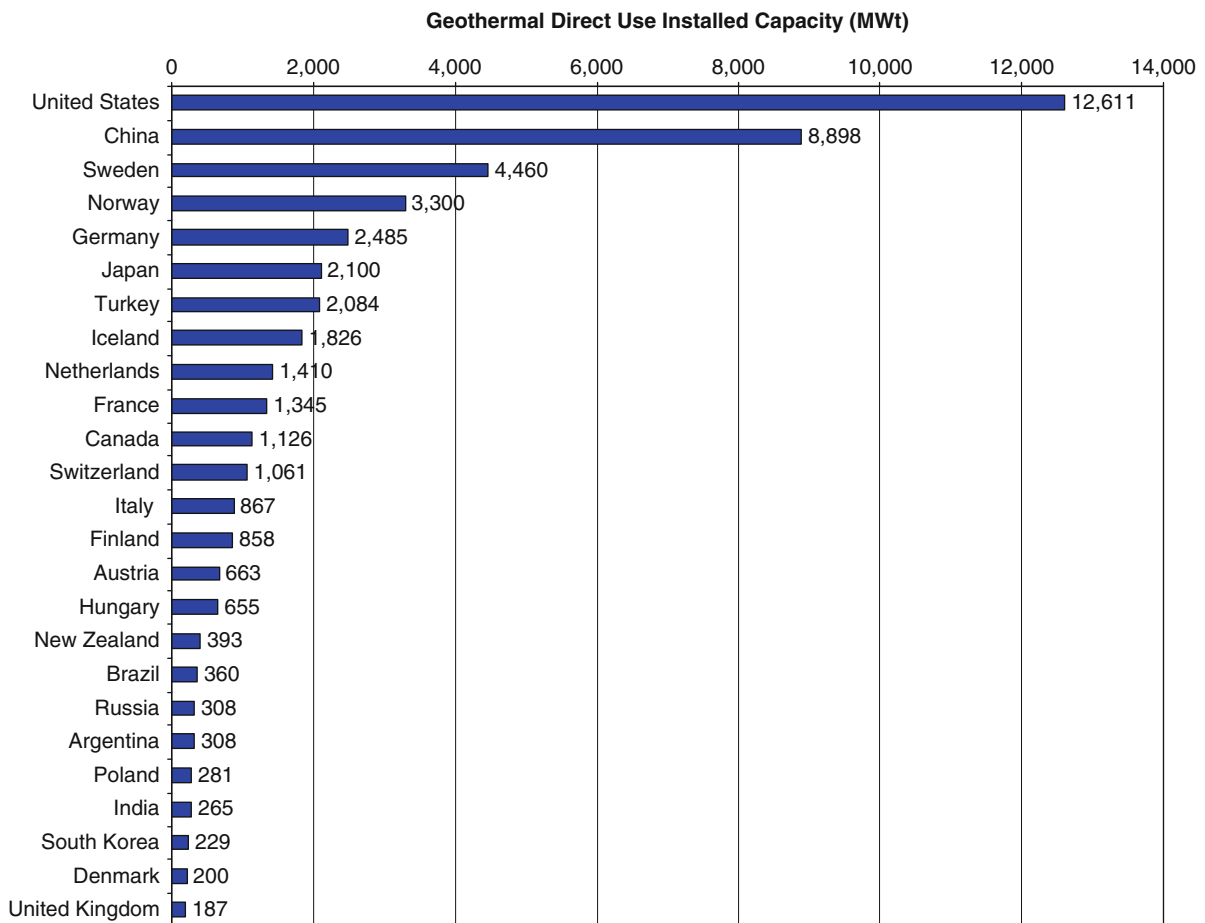
Geothermal Conditioning: Critical Sources for Sustainability. Figure 23
Global direct use geothermal application by percentage of total energy use 2010

conditioning with ground source heat pumps, 25% for bathing and swimming, 14.4% for direct space heating, and the remainder largely for greenhouses, aquaculture, industrial process heating, agricultural drying, space cooling, and snow melting. The annual energy savings from geothermal use based on 2010 data were equivalent to 307.8 million barrels of oil (Figs. 22 and 23).

Four countries accounted for almost 60% of the installed direct use capacity (Fig. 24): the United States (12,611 MWt), China (8,898 MWt), Sweden (4,460 MWt), and Norway (3,300 MWt). The largest increases in installed capacity between 2005 and 2010 occurred in the United Kingdom, South Korea, Ireland, Spain, and the Netherlands,

with heat pumps accounting for all capacity additions.

The countries with the greatest geothermal direct use energy consumption per year are China, the United States, Sweden, and Turkey, accounting for 49% of the annual total gigawatt-hours (see Table 6). When energy use per person is calculated, Iceland's geothermal direct use energy per person far exceeds all other countries (see Fig. 25). This results from the prevalence of direct use geothermal systems for space heating (such systems account for 89% of space heating in Iceland [56]) and to high energy content of Iceland's geothermal resources. "Low temperature" wells serving Reykjavik's district heating system, for example, supply water at 62–132°C [57]. Referring back to Fig. 2, the revised



Geothermal Conditioning: Critical Sources for Sustainability. Figure 24

Geothermal direct use installed capacity (MWt)

Geothermal Conditioning: Critical Sources for Sustainability. Table 6 Countries with highest geothermal direct use energy per year

Country	Annual use (GWh/year)
China	20,932
United States	15,710
Sweden	12,585
Turkey	10,247
Japan	7,139
Norway	7,001
Iceland	6,768
France	3,592
Germany	3,546
Netherlands	2,972
Italy	2,762
Hungary	2,713
New Zealand	2,654
Canada	2,465
Finland	2,325
Switzerland	2,143
Brazil	1,840
Russia	1,707
Mexico	1,118
Argentina	1,085
Austria	1,036
Slovak Republic	852
India	707
Denmark	695
Israel	609

Lindal diagram for low-exergy conditioning, it is interesting to note that this temperature range supports cogeneration with ample residual heat for conditioning and cascaded uses such as greenhouse heating.

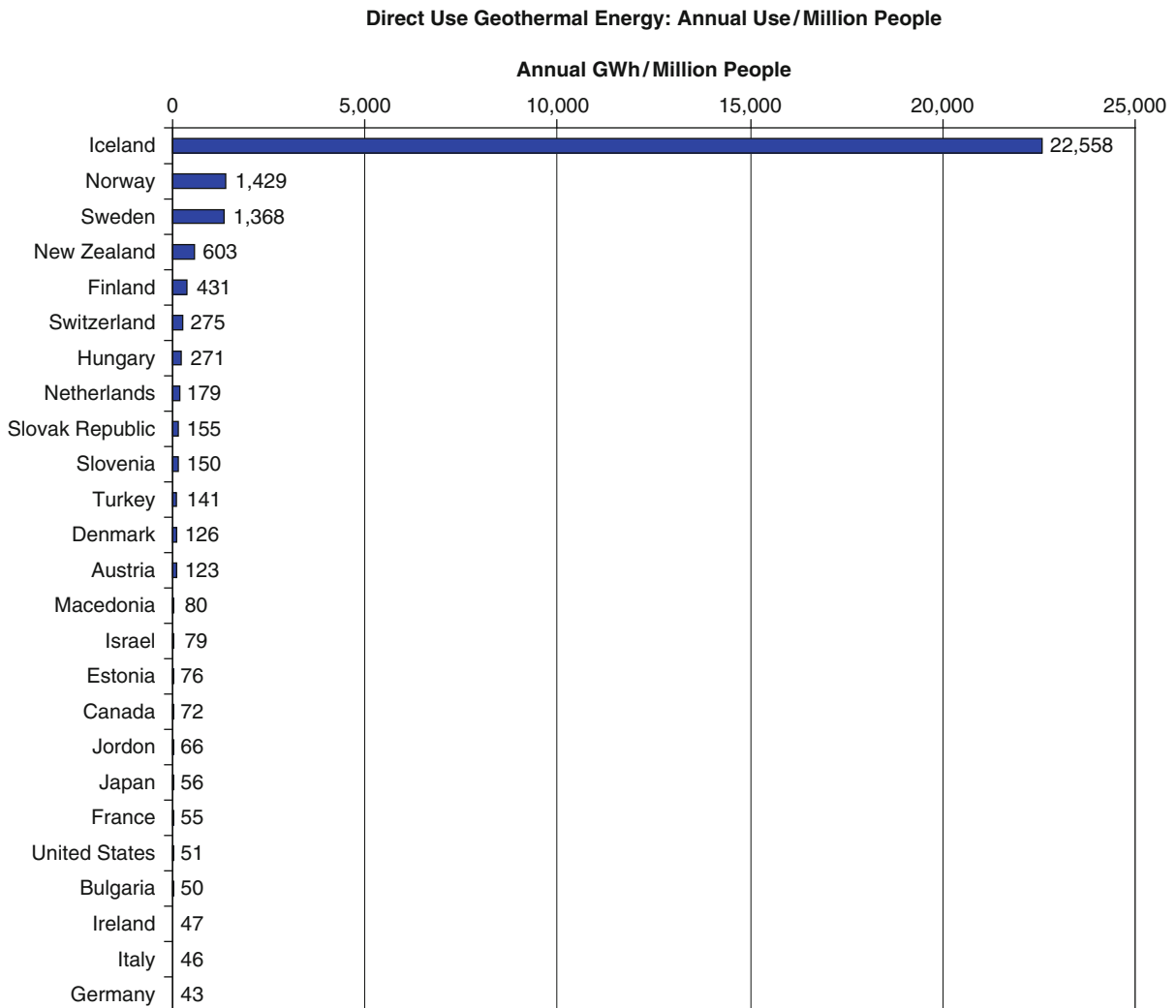
Future Directions

A study published by Ferguson and Woodbury in 2004 indicated that conductive heat loss from floors

and basements of buildings in Winnipeg, Canada was the likely source of a regional groundwater temperature anomaly beneath the city [59]. Just as buildings and their mechanical conditioning systems can generate urban heat islands above the earth's surface, they can also alter the geothermal gradient. Fortunately, a growing emphasis on well-designed building enclosures allows heating and cooling to be provided at temperatures much closer to the human comfort range and to cut extreme heat loss or gain at the perimeter. When occupant comfort requirements can be met with low-exergy systems, building conditioning operates in a range close to the ground temperature and buildings can be coupled with the substantial thermal energy supply and storage capacity the earth offers at the building foundation. The complementary development of high-performance enclosures and geothermal conditioning infrastructure – if accompanied by ongoing monitoring and research – has the potential to appreciably reduce building energy consumption, peak demand, and corresponding carbon emissions while maintaining occupant comfort and supporting environmental health.

To some extent, further development of geothermal conditioning infrastructure is limited by issues of cost, convenience, and property ownership. Nevertheless, there are several technical advances that could further increase its use and sustainability. These developments, some of which are underway, and others that are logical extensions of trends in sustainable building design and operation, may include:

- Increased efficiency of geothermal heat pumps
- Increased use of natural refrigerants such as R-744 (carbon dioxide)
- Improved understanding and modeling of soil mechanical behavior and temperature recovery time periods for ground heat exchangers
- New sustainable approaches to bacteriological and chemical fouling in heat exchangers, piping, and associated equipment
- Improvements in the cost, size, and efficiency of ground heat exchangers
- Increased use of heat recovery from existing underground infrastructure (e.g., sewer pipe, treated wastewater pipe) for building conditioning



Geothermal Conditioning: Critical Sources for Sustainability. Figure 25

Direct use geothermal energy: annual use/million people [58]

- Improvements in mechanical systems integration and system metering and control
- Increased use of district geothermal and solar/geothermal systems
- Development of vapor compression heat pumps powered by biogas motors or waterpower, or high-temperature gas absorption heat pumps powered by biomass [60, 61]

Expanded geothermal power production will support some of these developments since its scale

provides opportunities for district systems and energy cascades. One of the key organizations that conducts and tracks geothermal conditioning research is the International Energy Agency Heat Pump Programme (IEA HPP, www.heatpumpcentre.org). Founded in 1978, HPP current member countries are Austria, Canada, Finland, France, Italy, Germany, Japan, the Netherlands, Norway, South Korea, Sweden, Switzerland, and the United States. Through its web site, news and information about member country activities are available.

The Ground-Reach Project, an international effort supported by the European Commission to evaluate the use of ground source heat pumps in meeting Kyoto targets, maintains a database of 52 GSHP projects in 15 European countries at <http://www.groundreach.eu/>. In the future, eight demonstration projects from Mediterranean countries and 46 case studies from European projects being field-tested through 2012 will be added to the database. A central source for accessing these case studies and other projects related to geothermal energy research and development is the European Geothermal Energy Council (EGEC) website (<http://egec.info/>).

Within the United States, the Geo-Heat Center at Oregon Institute of Technology, hosts an online library that emphasizes applied (how-to) engineering of direct use systems <http://geoheat.oit.edu/publist.htm>. With publication dates ranging from 1975 to 2008, the system details and lessons learned are particularly valuable in this library.

Bibliography

Primary Literature

1. Lund J (2004) Geothermal direct use. In: Encyclopedia of energy. Elsevier, New York
2. Lund J (2004) Geothermal direct use. In: Encyclopedia of energy. Elsevier, New York, p 863
3. Dickson MH, Fanelli M (2004) What is geothermal energy? Geoscienze Istituto e Georisorse, CNR, Pisa. www.geothermal-energy.org/files-31.html
4. Aparicio E (2008) Urban surface water as energy source & collector. M.Sc. thesis, Delft University of Technology, Delft
5. Kaltschmitt M, Streicher W, Wiese A (2007) Basics of renewable energy supply. In: Renewable energy. Springer, Heidelberg/Berlin, p 92, 23–102
6. Clauser C (2009) Heat transport processes in the Earth's crust. *Surv Geophys* 30(3):163–191
7. After Table 2.6 in Kaltschmitt M, Streicher W, Wiese A (2007) Basics of renewable energy supply. In: Renewable energy: technology, economics and environment. Springer, Heidelberg/Berlin, pp 23–102
8. Short, Nicholas. "Geology, Weather and Climate: A Condensed Primer." http://rst.gsfc.nasa.gov/Sect9/Sect9_4.html
9. Oke TR (1987) Boundary layer climates, 2nd edn. London, Methuen, p 41
10. Kaltschmitt M, Streicher W, Wiese A (2007) Basics of renewable energy supply. In: Renewable energy. Springer, Heidelberg/Berlin, p 98, 23–102
11. Kavanaugh S, Rafferty K (1997) Ground source heat pumps: design of geothermal systems for commercial & institutional buildings. American Society of Heating Refrigerating and Air Conditioning, Atlanta
12. Figure 2.57, p 95, cited as Kaltschmitt M (2006/2007) Renewable energies: lessons. Institute for Environmental Technology and Energy Economics, Hamburg University of Technology, Hamburg
13. Clauser C (2006) Geothermal Energy. In: Heinloth K (ed) Landolt-Börnstein, Group VIII: advanced materials and technologies, vol 3, Energy technologies, Subvol. C: renewable energies. Springer, Heidelberg/Berlin, pp 493–604
14. Majorowicz J, Grasby S, Skinner W (2009) Estimation of shallow geothermal energy resource in Canada: heat gain and heat sink. *Nat Resour Res* 18(2):95–108
15. Björnsson S (2010) Geothermal development and research in Iceland. Orkustofnun and The National Energy Authority. http://www.nea.is/media/utgafa/GD_loka.pdf
16. Largely developed from work in building systems integration by Vivian Loftness, FAIA, and Volker Hartkopf, Dr. Ing., Dr.h.c., Center for Building Performance & Diagnostics, Department of Architecture, Carnegie Mellon University, Pittsburgh, PA (USA)
17. Quaschnig V (2010) Heat pumps – from cold to hot. In: Renewable energy and climate change. Wiley, Hoboken, pp 223–236
18. Laue HJ (2006) Heat pumps. In: Advanced materials and technologies, vol 3, Energy technologies, Subvol. C: renewable energy. Springer, Heidelberg/Berlin
19. American Society of Heating Refrigerating and Air-Conditioning Engineers (2002) Applied heat pump and heat recovery systems. In: ASHRAE handbook: HVAC systems and equipment. American Society of Heating, Refrigerating and Air Conditioning Engineers, Atlanta
20. Kavanaugh S, Rafferty K (1997) Heat pumps for ground source applications. In: American Society of Heating Refrigerating and Air Conditioning Engineers (ed) Ground source heat pumps: design of geothermal systems for commercial and institutional buildings. American Society of Heating Refrigerating and Air Conditioning Engineers, Atlanta
21. Kaltschmitt, Figure 9.1, p 389, cited as Halozan H, Holzapfel K (1987) Heizen mit Wärmepumpen, TÜV Rheinland, Köln
22. Harvey LDD (2006) A handbook on low-energy buildings and district-energy systems: fundamentals, techniques and examples. Earthscan, London. In North America, heat pump cooling efficiency is expressed as an Energy Efficiency Ratio (EER), which is an instantaneous, steady-state ratio of heat removed to the rate of energy used, in BTU/hr per watt
23. Brandl H (2006) Energy foundations and other thermo-active ground structures. *Geotechnique* 56(2):81–122
24. Nowak T (2009) Heat pumps are renewable – are they not? IEA Heat Pump Centre. <http://www.heatpumpcentre.org/en/newsletter/previous/Sidor/default.aspx>
25. American Society of Heating Refrigerating and Air-Conditioning Engineers (2007) ASHRAE handbook: HVAC applications, vol 32, Geothermal energy. American Society of Heating, Refrigerating and Air Conditioning Engineers, Atlanta, p 23

26. Kavanaugh S, Rafferty K (1997) Heat pumps for ground source applications. In: Ground source heat pumps: design of geothermal systems for commercial and institutional buildings. American Society of Heating Refrigerating and Air Conditioning Engineers, Atlanta. Reprint of Figure 1.6, Required design steps for GSHPs
27. American Society of Heating Refrigerating and Air Conditioning Engineers (2007) ASHRAE handbook: HVAC applications, vol 32, Geothermal energy. American Society of Heating, Refrigerating and Air Conditioning Engineers, Atlanta, pp 4–5
28. Kavanaugh S, Rafferty K (1997) Heat pumps for ground source applications. In: American Society of Heating Refrigerating and Air Conditioning Engineers (ed) Ground source heat pumps: design of geothermal systems for commercial and institutional buildings. American Society of Heating Refrigerating and Air Conditioning Engineers, Atlanta, p 22
29. Rafferty K (1998) Heat Exchangers. In: Lund JW, Lienau JP, Lunis B (eds) Geothermal direct use engineering and design guidebook, 3rd edn. Klamath Falls, Geo-Heat Center, Oregon Institute of Technology
30. Canadian GeoExchange Coalition (2010) Codes, standards and regulations in the Canadian GeoExchange industry: report of a national consultation conducted by the Canadian GeoExchange coalition (Summary), p 11. http://www.geo-exchange.ca/en/UserAttachments/news433_Standards%20Consultation%20-%20Final%20Report_Public%20_2010_E.pdf
31. American Society of Heating Refrigerating and Air Conditioning Engineers (2007) ASHRAE handbook: HVAC applications, vol 32, Geothermal energy. American Society of Heating, Refrigerating and Air Conditioning Engineers, Atlanta, p 13
32. Sanner DB (2007) Geothermal energy – opportunities for industry. In: EMEA environmental health and safety conference, Bruxelles, 13 June 2007. <http://www.egeg.org/target/bruxelles%20130607%20&j.pdf>
33. Kwok AG, Grondzik WT (2007) The green studio handbook: environmental strategies for schematic design. Architectural Press, Oxford, pp 169–174
34. Pfaffertott J, Walker-Hertkorn S, Sanner B (2007) Ground cooling: recent progress. In: Santamouris M (ed) Advances in passive cooling. Earthscan, London, pp 190–227
35. Lechner N (2009) Heating, cooling, lighting: sustainable design methods for architects, 3rd edn. Wiley, Hoboken, p 292
36. Nagano K (2009) GSHP in Japan. IEA Heat Pump Centre Newsletter 27(1). http://www.heatpumpcentre.org/en/newsletter/previous/Documents/HPC-news_1_2009.htm
37. Peron H (2010) Geotechnical design of heat exchanger piles. In: GSHP association research seminar, current and future research into ground source energy. National Energy Centre, Milton Keynes, United Kingdom. <http://www.gshp.org.uk/documents/REsearchseminar2010/Herve%20Peron%20Geotechnical%20Design%20of%20Heat%20Ex.pdf>
38. Clauser C (2006) Geothermal Energy. In: Heinloth K (ed) Landolt-Börnstein, Group VIII: advanced materials and technologies, vol 3, Energy technologies, Subvol. C: renewable energies. Springer, Heidelberg/Berlin, pp 493–604, Figure 8.31
39. Radial systems are mentioned in Kaltschmitt M (2007) Utilisation of ambient air and shallow geothermal energy. In: Kaltschmitt M, Wolfgang S, Andreas W (eds) Renewable energy: technology, economics and environment. Springer, Heidelberg/Berlin, pp 385–436
40. Baird N (2009) Center for Building Performance & Diagnostics. School of Architecture, Carnegie Mellon University
41. Kaltschmitt M (2007) Utilisation of geothermal energy. In: Kaltschmitt M, Wolfgang S, Andreas W (eds) Renewable energy: technology, economics and environment. Springer, Heidelberg/Berlin, p 465
42. Kaltschmitt M (2007) Utilisation of geothermal energy. In: Kaltschmitt M, Wolfgang S, Andreas W (eds) Renewable energy: technology, economics and environment. Springer, Heidelberg/Berlin, p 463
43. Paksoy, J., A. Snijders and L. Stiles. “Aquifer Thermal Energy Cold Storage System at Richard Stockton College,” 2009. http://talon.stockton.edu/eyos/energy_studies/content/docs/effstock09/Session_6_3_ATES_Applications/57.pdf
44. Drake Landing Solar Community. <http://www.dlsc.ca/>
45. Driscoll FG (1986) Groundwater and wells. St. Paul, Johnson Filtration Systems, p 1, 2 Sub edn
46. Baird N (in progress) Critical guidelines for energy and water performance of open loop geothermal systems for low energy building conditioning. Doctoral dissertation, School of Architecture, Carnegie Mellon University, Pittsburgh
47. Baird N (2008) Center for Building Performance & Diagnostics. School of Architecture, Carnegie Mellon University
48. Clauser C (2006) Geothermal Energy. In: Heinloth K (ed) Landolt-Börnstein, Group VIII: advanced materials and technologies, vol 3, Energy technologies, Subvol. C: renewable energies. Springer, Heidelberg/Berlin, pp 493–604, Figure 8.33
49. Baird N (2010) Center for Building Performance & Diagnostics. School of Architecture, Carnegie Mellon University
50. NL EVD International (2005) China: cold rise (PESP01051). Publication 147970. <http://www.evd.nl/cooperation/zoeken/showbouwsteen.asp?bstnum=147970&location=&highlight=>
51. Harvey LDD (2006) A handbook on low-energy buildings and district-energy systems: fundamentals, techniques and examples. Earthscan, London, p 587. In North America, heat pump cooling efficiency is expressed as an Energy Efficiency Ratio (EER), which is an instantaneous, steady-state ratio of heat removed to the rate of energy used, in BTU/hr per watt
52. Rabtherm – Energy Systems AG. <http://www.rabtherm.com/>
53. Sanner B, Kabus F, Seibt P, Bartels J (2005) Underground thermal energy storage for the German parliament in Berlin, system concept and operational experience. In: World geothermal congress 2005, Antalya

54. Roijen E, Op't Veld P, Demollin-Schneiders E (2007) The mine water project Heerlen – low exergy heating and cooling in practice. In: PALENC AIVC 2007. <http://www.chri.nl/upload/art.%20minewaterproject.pdf>
55. Municipality Heerlen. <http://www.heerlen.nl/Pub/Duurzaamheid/Projecten-Stadsplanning-Mijnwaterenergie.html>
56. Lund J, Freeston D, Boyd T (2010) Direct utilization of geothermal energy 2010 worldwide review. In: World geothermal congress, Bali
57. Björnsson S (2010) Geothermal development and research in Iceland. Orkustofnun and The National Energy Authority, p 21. http://www.nea.is/media/utgafa/GD_loka.pdf
58. This table was generated using the Lund et al (2010) World-wide review of geothermal energy use. See Lund J, Freeston D, Boyd T (2010) Direct utilization of geothermal energy 2010 worldwide review. In: World geothermal congress, Bali and population data from http://en.wikipedia.org/wiki/List_of_countries_by_population and http://en.wikipedia.org/wiki/List_of_Caribbean_island_countries_by_population
59. Ferguson G, Woodbury AD (2004) Subsurface heat flow in an urban environment. *J Geophys Res* 109:B02402
60. Banks D (2008) An introduction to thermogeology: ground source heating and cooling. Wiley-Blackwell, Oxford, p 296
61. European Geothermal Energy Council (2011) Strategic Research Agenda: Geothermal Heating and Cooling. <http://egec.info/>

Books and Reviews

- Carmody J, Sterling R (1983) *Underground building design: commercial and institutional structures*. Van Nostrand Reinhold, New York
- Dickinson JS, Buik N, Matthews MC, Snijders A (2009) Aquifer thermal energy storage: theoretical and operational analysis. *Geotechnique* 59(3):249–260
- Givoni B (1998) *Climate considerations in building and urban design*. Van Nostrand Reinhold, New York
- International Energy Agency (2011) *Technology Roadmap: Geothermal Heat and Power*. www.iea.org
- Lienau P (1998) *Geothermal direct use engineering and design guidebook*, 3rd edn. Oregon Institute of Technology, Geo-Heat Center, Klamath Falls
- Orio CD, Chlasson A, Johnson CN, Deng Z, Rees SJ, Spitler JD (2005) A survey of standing column well installations in North America. *Trans Am Soc Heat Refrig Air-Cond Eng (ASHRAE)* 111(2):109–121
- Powrie W, Preene M (2009) Ground energy systems: from analysis to geotechnical design. *Géotechnique* 59(3):261–271
- Wood CJ, Liu H, Riffat SB (2009) Use of energy piles in a residential building, and effects on ground temperature and heat pump efficiency. *Geotechnique* 59(3):287–290
- Xu X et al (2010) Active pipe-embedded structures in buildings for utilizing low-grade energy sources: a review. *Energy Build* 42(10):1567–1581

Geothermal Energy Utilization

JOHN W. LUND

Geo-Heat Center, Oregon Institute of Technology,
Klamath Falls, OR, USA

Article Outline

Glossary
 Definition and Importance of Geothermal Energy
 Introduction
 Types of Geothermal Resources
 Utilization in 2010
 Environmental Considerations
 Energy Savings
 Future Directions
 Bibliography

Glossary

Balneology The science of the healing qualities of baths, especially with natural mineral waters; the therapeutic use of natural, warm, or mineral waters.

Binary power plant Used with low-temperature resources (below 150°C or 300°F) where a secondary low boiling point working fluid (normally a hydrocarbon) is vaporized by the geothermal fluid through a heat exchanger to drive a turbine producing electricity. Also referred to as an organic Rankine cycle (ORC) machine.

Caldera A large basin-shaped volcanic depression, circular in form, with a diameter many times greater than the included volcanic vent usually causes by an explosive volcanic eruption that drains the magma chamber resulting in collapse of the volcano.

Calorie The quantity of heat needed to raise the temperature of 1 gram (g) of water by 1 degree centigrade (°C) at 16°C. It is equal to 4.185 J.

Cap rock A comparatively impervious stratum that prevents the circulation of heat or fluid.

Conduction The transfer of heat through a medium or body driven by a temperature gradient and involving no particle motion. The average temperature gradient of the world, caused by conduction is about 25°C/km increasing with depth above the mean annual surface temperature.

Convection A process of mass movements of portions of any fluid medium (liquid or gas) as a consequence of different temperatures in the medium and hence different densities moving the medium and also the heat.

Enhanced (engineered) geothermal systems (EGS) Extracting heat stored in rocks within about 10 km of the surface, from which energy cannot be economically extracted by natural hot water or steam. The system is hydrofractured and water pumped down one well, extracting the heat by flowing through the fractures, and producing hot water or steam through a second well.

Fault A fracture or fracture zone along which there has been displacement of the sides relative to one another parallel to the fracture. The movement can be vertical, horizontal, or a combination of the two.

Flash steam The steam generated when the pressure on hot water (usually above 100°C) is reduced.

Fossil fuel A deposit of organic material containing stored solar energy that can be used as fuel, such as coal, natural gas, and petroleum.

Fumarole A hole or vent from which fumes or vapors issue usually found in volcanic areas.

Geopressured Zones below depths of 1,800–3,000 m, in which sediments in basins are commonly characterized by abnormally high pressure, high temperature, and high salinity.

Geothermal energy The internal energy of the earth, usually from the radioactive decay of potassium, thorium, and uranium, often associated with magma bodies, available to humans as heat from heated rocks, water, or steam.

Geyser A spring that erupts with intermittent jets of heated water or steam.

Heat exchanger A device for transferring heat from one fluid to another. The fluids are usually separated by conducting walls of metal or plastic.

Heat flow Dissipation of heat coming from within the earth by conduction. The worldwide average is about 65 mW/m².

Heat pump A device which, by the consumption of work or heat, affects the transport of heat between a lower temperature to a higher temperature source. The useful output is heat in conventional

usage. The reverse process is called a refrigerator used for the removal of heat.

Hot Spring A thermal spring whose water has a higher temperature than that of the human body (usually above 40°C).

Hydrothermal An adjective applied to heated or hot aqueous-rich solutions, to the processes of which they are concerned, and to the rocks, ore deposits, and alteration products produced by them.

Joule (J) The SI unit for all forms of energy or work. It is equal to 1 W-s or 0.239 cal.

Lava Hot fluid rock that issues from a volcano or a fissure in the earth's surface coming from subsurface magma.

Magma Molten rock within the earth from which an igneous rock results by cooling, and forms lava when it erupts on the earth's surface.

Permeability The capacity of a rock to transmit fluid, dependent upon the size and shape of the pores and their interconnections.

Seismic Pertaining to an earthquake or earth vibrations, including those that are artificially induced.

Spa A resort using mineral water for bathing, soaking, and drinking along with covering portions of the body with mineral muds for therapeutical purposes. Diet, exercise, and rest can also be part of the spa treatment plan.

Subsidence A sinking of a large part of the earth's crusts, often due to the removal of fluid by pumping.

Volcano A vent in the earth surface through which magma as lava and associated gases, and/or pyroclastic material (rock, cinders, pumice, and ash) erupt.

Watt (W) A unit of power or energy produced over time, equivalent to 1 J/s, or 0.001341 horse power (hp).

Definition and Importance of Geothermal Energy

Geothermal energy is the heat contained within the Earth that generates geological phenomena on a planetary scale. The main sources of this energy are due to the heat flow from the earth's core and mantle generated by the radioactive decay of potassium, thorium, and uranium in the crust or by friction heat generated in subduction zones along continental plate

margins. It may be characterized by surface expression of fumaroles, hot springs, geysers, volcanic eruption, and lava flows. Geothermal energy is often used to indicate that part of the Earth's heat that can, or could, be recovered and exploited by humankind. The resource is large, is renewable in the broad sense, and is available almost everywhere in the world, depending upon the depth to the resource and the economics to produce it. The total estimated thermal energy above surface temperature to a depth of 10 km under the continents, reachable with current drilling technology, is 1.3×10^{27} J (1.3×10^9 EJ = exajoules). Recovery of geothermal energy utilizes only a portion of the stored thermal energy due to limitations in rock permeability that permit heat extraction through fluid circulation, and to the minimum temperature limits for utilization at a given site. The recovery factor is estimated between 0.5% and 20% [1]; and at the lower rate, this is 6.5×10^6 EJ, or about 200,000 TW-years (terawatt = 10^{12} W). This is about three times the annual world consumption for all types of energy, and about 130 times at the higher recovery rate.

Geothermal energy can be used over a range of temperature to supply electricity, and heat and cool for the benefit of humankind. The higher temperature (above 175°C) is traditionally used to produce electricity; however with the improvement in the organic Rankine cycle or binary power plants described later, the usable temperature has been reduced to around 100°C. Lower temperatures are used for direct heating and cooling, from industrial process heating, space heating, and cooling including district energy systems, the heating of greenhouses and aquaculture ponds, to heating swimming pools and spas, generally in the range of 40–150°C. Finally the lowest temperatures from 5°C to 30°C, available anywhere in the world at shallow depth (up to 300 m), can utilize geothermal heat pumps for space heating and cooling.

Introduction

Early humans probably used geothermal water that occurred in natural pools and hot springs for cooking, bathing, and to keep warm [2]. There is archeological evidence that the Indians of the Americas occupied sites around these geothermal resources for over 10,000 years to recuperate from battle and take refuge.

Many of their oral legends describe these places and other volcanic phenomena. Recorded history shows uses by Romans, Japanese, Turks, Icelanders, Central Europeans, and the Maori of New Zealand for bathing, cooking, and space heating. Baths in the Roman Empire, the middle kingdom of the Chinese, and the Turkish baths of the Ottomans were some of the early uses of balneology, where body health, hygiene, and discussions were the social custom of the day. This custom has been extended to geothermal spas in Japan, Germany, Iceland, and countries of the former Austro-Hungarian Empire, the Americas, and New Zealand. Early industrial applications include chemical extraction from the natural manifestations of steam, pools, and mineral deposits in the Larderello region of Italy, with boric acid being extracted commercially starting in the late 1700s. At Chaudes-Aigues in the heart of France, the world's first geothermal district heating system was started in the fourteenth century and is still going strong. The oldest and still operating geothermal district heating project in the United States is on Warm Springs Avenue in Boise, Idaho, going on line in 1892 and providing space heating for up to 450 homes.

The first use of geothermal energy for electric power production started in Italy with experimental work by Prince Gionori Conti between 1904 and 1905. The first commercial power plant (250 kWe) was commissioned in 1913 at Larderello, Italy. These developments were followed by flash steam plants coming on line in New Zealand at Wairakei in 1958; an experimental plant at Pathe, Mexico, in 1959; and the first commercial plant at The Geysers in the United States in 1960. Japan followed with 23 MWe at Matsukawa in 1966. All of these early plants used steam directly from the earth (dry-steam fields), except for New Zealand, which was the first to use flashed or separated steam for running the turbines. The former USSR produced power from the first true binary power plant, 680 kWe using 81°C water at Paratunka on the Kamchatka peninsula – the lowest temperature ever reported used in the world for power generation from geothermal energy at that time. Iceland first produced power at Namafjall in northern Iceland, from a 3 MWe noncondensing turbine. These were followed by plants in El Salvador, China, Indonesia, Kenya, Turkey, Philippines, Portugal (Azores), Greece, and Nicaragua in the 1970s and 1980s. Later

plants were installed in Thailand, Argentina, Taiwan, Australia, Costa Rica, Austria, Guatemala, Ethiopia, with the latest installations in Germany and Papua New Guinea. Recently in 2006, a 200 kW binary plant was started at Chena Hot Springs in Alaska using geothermal fluids at 74°C, the lowest temperature for electric power generation recorded to date [3].

Types of Geothermal Resources

Geothermal energy comes from the natural heat of the earth primarily due to the decay of the naturally radioactive isotopes of uranium, thorium, and potassium. Because of the internal heat, the Earth's surface heat flow averages 65 mW/m² which amounts to a total heat loss of about 44 million megawatts (1,400 EJ/year). The estimated total thermal energy above surface temperature to a depth of 10 km, the limit of the deepest exploration drilling, is 1.3×10^{27} J (1.3×10^9 EJ), equivalent to burning 3.0×10^{17} barrels of oil. Since the global energy consumptions for all types of energy is equivalent to the use of about 100 million barrels of oil per day, the Earth's energy to a depth of 10 km would supply all of humankind's energy needs for six million years [4].

On average, the temperature of the Earth with depth increases about 25°C/km above the surface ambient temperature. Thus, assuming a conductive gradient, the temperature of the earth at 10 km would be over 250°C. However, most geothermal exploration and use occurs where the gradient is higher, and thus where drilling is shallower and less costly. These shallow depth geothermal resources occur due to: (1) intrusion of molten rock (magma) from depth, bringing up great quantities of heat; (2) high surface heat flow, due to a thin crust and high-temperature gradient; (3) ascent of groundwater that has circulated to depths of several kilometers and been heated due to the normal temperature gradient; (4) thermal blanketing or insulation of deep rocks by thick formation of such rocks as shale whose thermal conductivity is low; and (5) anomalous heating of shallow rock by decay of radioactive elements, perhaps augmented by thermal blanketing [4].

Geothermal resources are usually classified as shown in Table 1, modeled after White and Williams [5]. These geothermal resources range from the mean annual ambient temperature of around 20°C to over

300°C. In general, resources above 150°C are used for electric power generation, although power has recently been generated at Chena Hot Springs Resort in Alaska using a 74°C geothermal resource [3]. Resources below 150°C are usually used in direct-use projects for heating and cooling. Ambient temperatures in the 5–30°C range can be used with geothermal (ground-source) heat pumps which provide both heating and cooling.

Convective hydrothermal resources occur where the Earth's heat is carried upward by convective circulation of naturally occurring hot water or steam. Underlying some high-temperature convective hydrothermal resources are temperatures of 500–1,000°C from molten intrusions of recently solidified rocks. The lower temperature resource results from deep circulation of water along fractures. *Vapor-dominated systems* (Fig. 1) produce steam from boiling of deep, saline waters in low permeability rocks. These reservoirs are few in number, with The Geysers in northern California, Larderello in Italy, and Matsukawa in Japan being ones where the steam is exploited to produce electric energy. *Water-dominated systems* (Fig. 2) are produced by ground water circulating to depth and ascending from buoyancy in permeable reservoirs that are a uniform temperature over large volumes. There is typically an upflow zone at the center of each convection cell, an outflow zone or plume of heated water moving laterally away from the center of the system, and a downflow zone where recharge is taking place. Surface manifestations include hot springs, fumaroles, geysers, travertine deposits, chemically altered rocks, or sometimes, no surface manifestations (a blind resource).

Sedimentary basins (Fig. 3) produce higher temperature resources than the surrounding formations due to their low thermal conductivity or high heat flow or both producing geothermal gradients >30°C/km. These generally extend over large areas and are typical of the Madison Formation of North Dakota, South Dakota, Montana, and Wyoming area of the northern United States and the Pannonian Basin of Central Europe where it has been used extensively in Hungary.

Geopressured resources (Fig. 4) occur in basin environments where deeply buried fluids contained in permeable sedimentary rocks warmed in a normal or enhanced geothermal gradient by their great burial

depth. The fluids are tightly confined by surrounding impermeable rock and bear pressure much greater than hydrostatic. Thermal waters under high pressure in sand aquifers are the target for drilling, mainly as they contain dissolved methane. The source of energy available from this type of resources consists of: (1) heat, (2) mechanical energy, and (3) methane. The Texas and Louisiana Gulf Coast in the United States has been

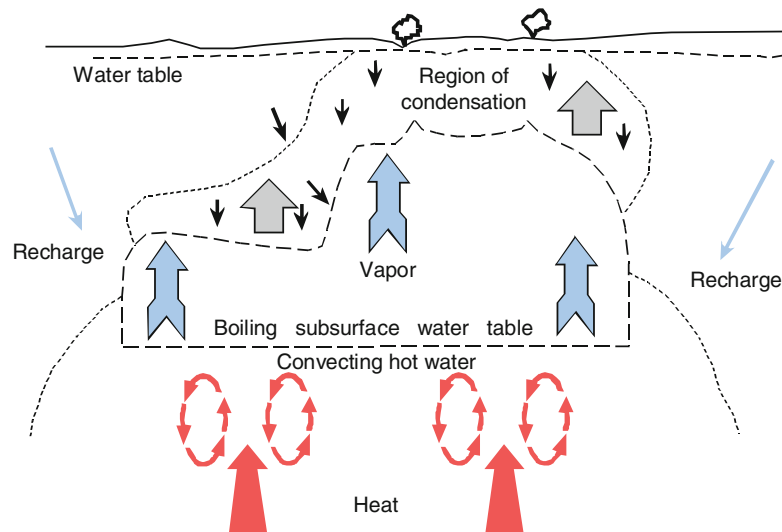
tested for the geothermal energy; however, due to the great depths of several kilometers, they have not proved economic, but are currently being evaluated again.

Radiogenic resources (Fig. 5) are found where granitic intrusions are near surface heating up the local groundwater from the decay of radioactive thorium, potassium, and uranium. This localized heating increases the normal geothermal gradient providing hot water at economical drilling depths. This type of resource occurs along the eastern United States, but has not been developed commercially.

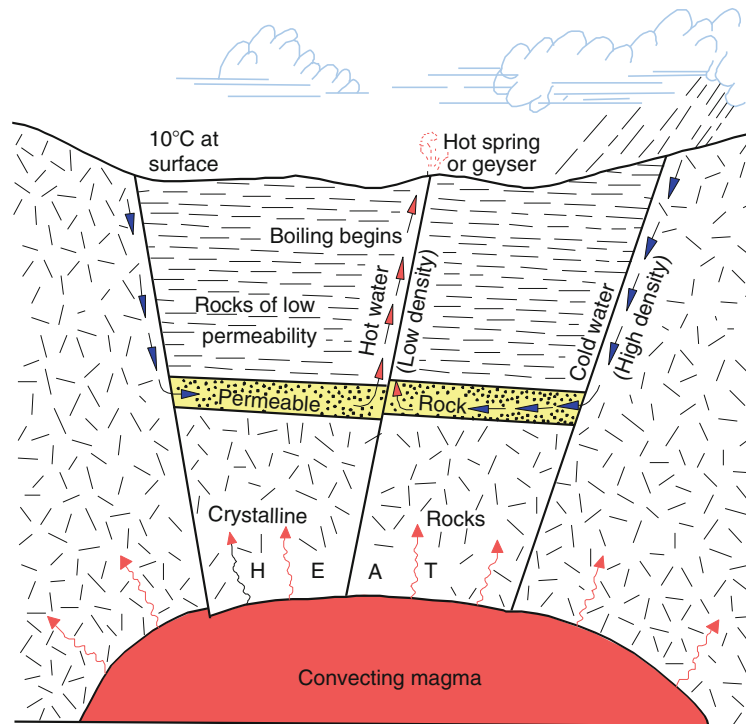
Hot dry rock resources (Fig. 6) are defined as heat stored in rocks within about 10 km of the surface from which energy cannot be economically extracted by natural hot water or steam. These hot rocks have few pore space or fractures, and therefore, contain little water and little or no interconnected permeability. In order to extract the heat, experimental projects have artificially fractured the rock by hydraulic pressure, followed by circulating cold water down one well to extract the heat from the rocks and then producing from a second well in a closed system. Early experimental projects were undertaken at Fenton Hill (Valdes Caldera) in northern New Mexico and on Cornwall in southwest England; however, both of these projects have been abandoned due to lack of funds and poor results. Projects are currently underway in Soultz-sous-Forêt in the Rhine

Geothermal Energy Utilization. Table 1 Geothermal resource types

Resource type	Temperature range (°C)
Convective hydrothermal resources	
Vapor dominated	≈240°
Hot water dominated	20° to 350°+
Other hydrothermal resources	
Sedimentary basin	20° to 150°
Geopressured	90° to 200°
Radiogenic	30° to 150°
Hot rock resources	
Solidified (hot dry rock)	90° to 650°
Part still molten (magma)	>600°



Geothermal Energy Utilization. Figure 1
Vapor-dominated geothermal system



Geothermal Energy Utilization. Figure 2
Hot water-dominated geothermal system

Graben on the French–German border, in Germany at Bad Urach, several locations in Japan, and in the Cooper Basin of Australia [6]. Renewed interest has been generated in the United States for enhanced (engineered) geothermal systems (EGS) based on a recent MIT report [7].

Molten rock or magma resources have been drilled in Hawaii experimentally to extract heat energy directly from molten rock. It has been used successfully at Heimaey in Iceland (one of the Westmann Islands) after the 1973 eruption. A heat exchanger constructed on the surface of the lava flow recovered steam resulting from boiling of downward percolation water from the surface. The heat was used in a space-heating system for over 10 years, but is now shutdown due to cooling of the surrounding rock.

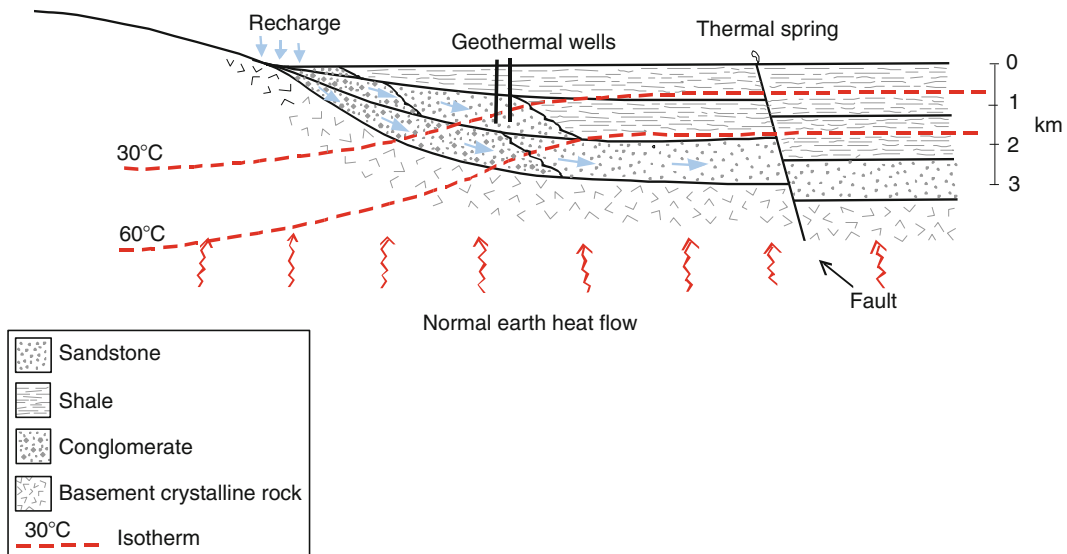
Utilization in 2010

Based on 68 country update papers submitted to the World Geothermal Congress 2010 (WGC2010) held in Bali, Indonesia, the following figures on

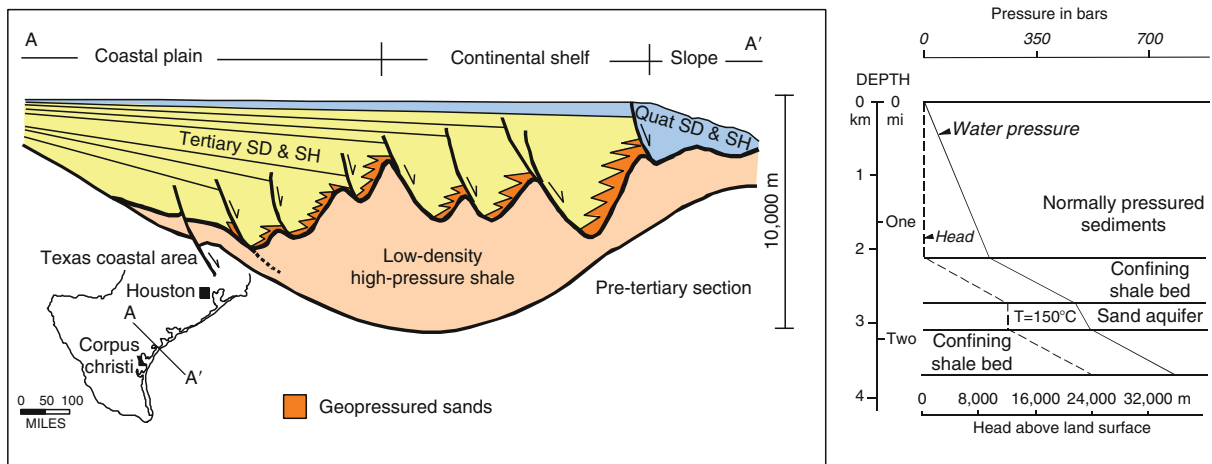
worldwide geothermal electric and direct-use capacity are reported. A total of 78 countries have reported some utilization from WGC2000, WGC2005, and WGC2010 electric, direct use, or both [8–12] (Table 2).

The figures for electric power capacity (MWe) appear to be fairly accurate; however, several of the country's annual generation values (GWh) had to be estimated which amounted to only 0.5% of the total. The direct-use figures are less reliable and probably are understated by as much as 20%. The author is also aware of at least five countries, which utilize geothermal energy for direct-heat applications, but did not submit reports to WGC2010. The details of the present installed electric power capacity and generation, and direct use of geothermal energy can be found in Bertani [12] and Lund et al. [10]. These data are summarized in Table 3.

A review of the above data show that for electric power generation each major continent has approximately the same percentage share of the installed capacity and energy produced with the Americas



Geothermal Energy Utilization. Figure 3
Sedimentary basin geothermal resource



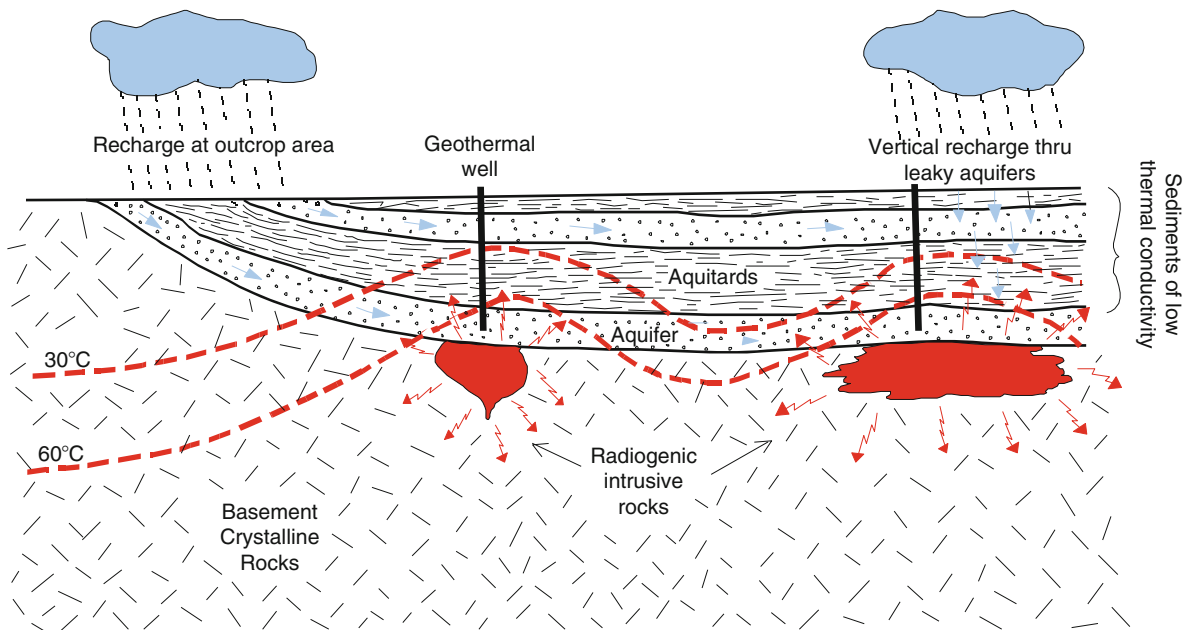
Geothermal Energy Utilization. Figure 4
Geopressed geothermal system

and Asia having over 75% of the total; whereas with the direct-use figures, the percentages drop significantly from installed capacity to energy use for the Americas (28.9–18.4%) due to the high percentage of geothermal heat pumps with low capacity factor for these units in the United States and Canada. On the other hand, the percentages are approximately equal for the remainder of the world

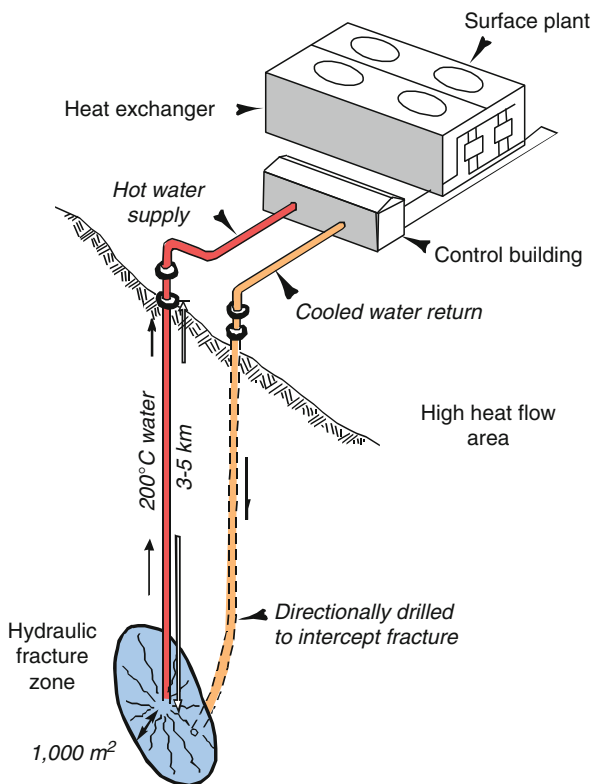
due to a lesser reliance on geothermal heat pumps, and the greater number of operating hours per year for these units.

Electric Power Generation

Geothermal power is generated by using steam or a hydrocarbon vapor to turn a turbine-generator set



Geothermal Energy Utilization. Figure 5
Radiogenic geothermal system



Geothermal Energy Utilization. Figure 6
Hot dry rock exploitation

to produce electricity. A vapor-dominated (dry steam) resource (see Figs. 1 and 7) can be used directly, whereas a hot water resource (see Figs. 2 and 8) needs to be flashed by reducing the pressure to produce steam. In the case of low-temperature resource, generally below 150°C, they require the use of a secondary low boiling point fluid (typically a hydrocarbon) to generate the vapor, in a binary or organic Rankine cycle plant (see Fig. 9). Usually a wet or dry cooling tower is used to condense the vapor after it leaves the turbine to maximize the temperature and pressure drop between the incoming and outgoing vapor and thus increase the efficiency of the operation. The worldwide installed capacity has the following distribution: 27% dry steam, 41% single flash, 22% double flash, 12% binary/combined cycle/hybrid, and 1% backpressure [12].

Electric power has been produced from geothermal energy in 27 countries; however, Greece, Taiwan, and Argentina have shut down their plants due to environmental and economic reasons. Since 2000, the installed capacity in the world has increased almost 3,000 MWe. Since 2000, additional plants have been installed in Costa Rica, France on Guadeloupe in the Caribbean, Iceland, Indonesia, Kenya, Mexico, and Philippines.

In 2004, Germany installed a 210-kWe binary plant at Neustadt Glewe and 56-MWe plants have been installed on Papua New Guinea to generate electricity for a remote mine. Russia has completed a new 50-MWe plant on Kamchatka. More recently, a 200 kW binary plant using 74°C geothermal water and 4°C cooling was installed at Chena Hot Springs Resort in Alaska [3]. The operating capacity in the United States has increased since 1995 due to completion of the two effluent pipelines injecting treated sewage water at The Geysers. In an attempt to bring production back, the Southeast Geysers Effluent Recycling Project is now injecting 340 l/s of treated wastewater through a 48-km long pipeline from Clear Lake, adding 77 MWe. A second, 66-km long pipeline from Santa Rosa was placed on line in 2004, injecting 480 l/s that are projected to add another 100 MWe to The Geysers' capacity. Table 4 lists the leading countries producing electric power.

One of the more significant aspects of geothermal power development is the size of its contribution to national and regional capacity and production of

countries. The following countries or regions (Table 5) lead in this contribution with more than 5% of the electrical energy supplied by geothermal power based in data from country update papers from WGC2010 [12].

Direct Utilization

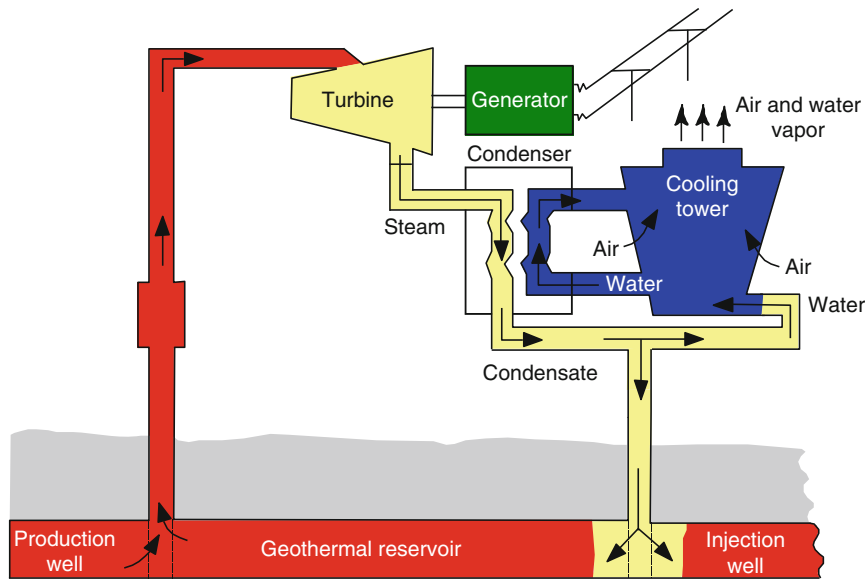
Direct use of geothermal resources is primarily for direct heat and cooling. The main utilization categories are: (1) swimming, bathing, and balneology; (2) space heating and cooling including district energy systems; (3) agricultural applications such as greenhouse and soil heating; (4) aquaculture application such as pond and raceway water heating; (5) industrial applications such as mineral extraction, food, and grain drying; and (6) geothermal (ground-source) heat pumps, used for both heating and cooling. Direct use of geothermal resources normally uses temperatures below 150°C as illustrated in Fig. 10. The main advantage of using geothermal energy for direct-use projects in this low-to intermediate-temperature range is that these resources are more widespread and exist in at least 80 countries at economic drilling depths. In addition, there are no conversion efficiency losses and projects can use conventional water-well drilling and off-the-shelf heating and cooling equipment (allowing for the temperature and chemistry of the fluid). Most projects can be on line in less than a year. Projects can be on a small scale ("mom and pop operations") such as for an individual home, single greenhouse, or aquaculture pond, but can also be a large-scale operation such as for district heating/cooling, for food and lumber drying, and mineral ore extraction.

Geothermal Energy Utilization. Table 2 Total geothermal capacity and use in 2010

Installed annual				
Use	Power (MW)	Energy use (GWh/year)	Capacity factor	Countries reporting
Electric power	10,715	67,246	0.72	24
Direct use	50,583	121,696	0.27	78

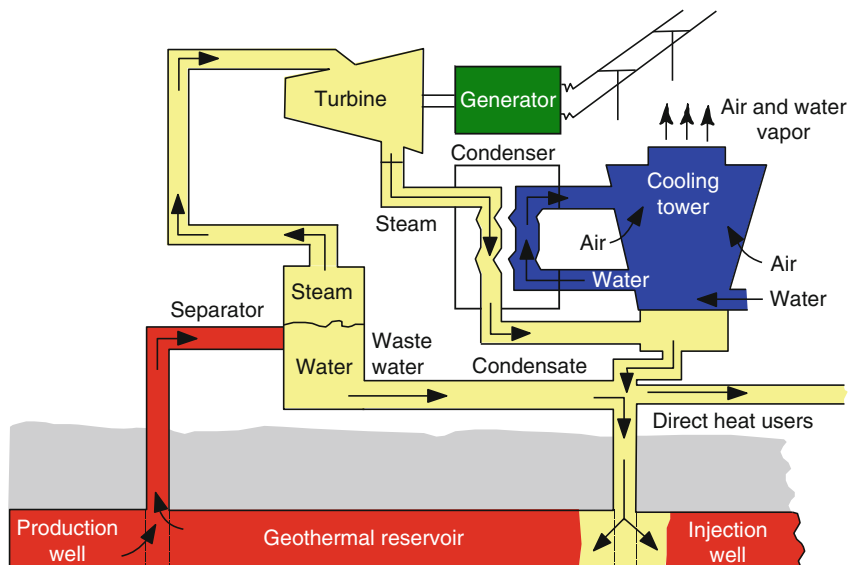
Geothermal Energy Utilization. Table 3 Summary of regional geothermal use in 2010

Electric power				Direct use		
Region	%MWe	%GWh/year	#countries	%MWt	%GWh/year	#countries
Africa	1.6	2.1	2	0.1	0.6	7
Americas	42.6	39.9	6	28.9	18.4	15
Asia	34.9	35.1	6	27.5	33.8	16
Europe	14.5	16.2	7	42.5	45.0	37
Oceania	6.4	6.7	3	1.0	2.2	3



Geothermal Energy Utilization. Figure 7

Steam plant using a vapor- or dry-steam-dominated geothermal resource

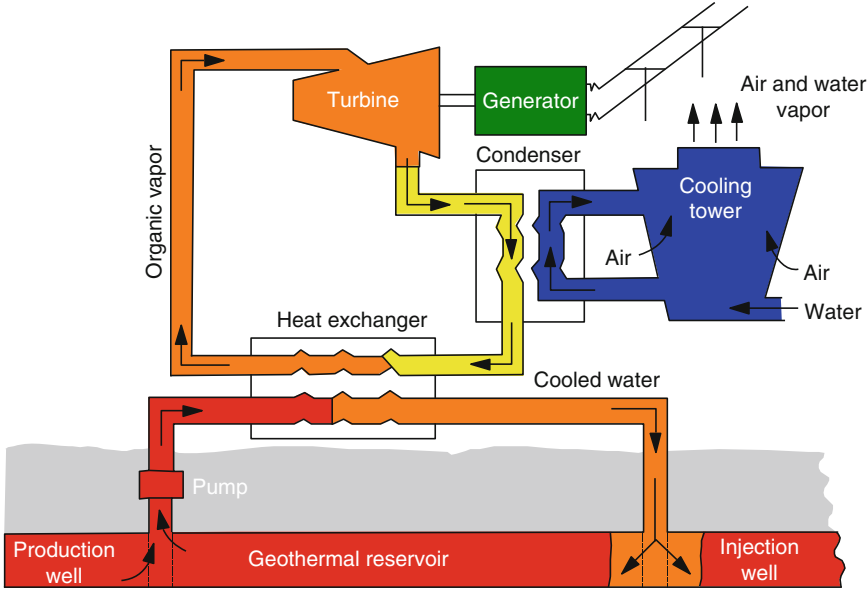


Geothermal Energy Utilization. Figure 8

Flash steam plant using a water-dominated geothermal resource with a separator to produce steam

It is often necessary to isolate the geothermal fluid from the user side to prevent corrosion and scaling. Care must be taken to prevent oxygen from entering the system (geothermal water normally is oxygen

free), and dissolved gases and minerals such as boron, arsenic, and hydrogen sulfide must be removed or isolated as they are harmful to plants and animals. On the other hand carbon dioxide, which often occurs



Geothermal Energy Utilization. Figure 9

Binary power or organic Rankin cycle plant using a low-temperature geothermal resource and a secondary fluid of a low boiling point (typically a hydrocarbon)

Geothermal Energy Utilization. Table 4 Leading countries in electric power generation (>100 MWe) [12]

Country	Installed capacity (MWe)	Running capacity ^a (MWe)	Annual energy produced (GWh/year)	Running capacity factor	Number of units operating
United States	3,093	2,024	16,603	0.94	209
Philippines	1,904	1,774	10,311	0.66	56
Indonesia	1,197	1,197	9,600	0.92	22
Mexico	958	958	7,047	0.84	37
Italy	843	843	5,520	0.75	33
New Zealand	628	628	4,055	0.74	43
Iceland	575	575	4,597	0.91	25
Japan	536	422	3,064	0.83	20
El Salvador	204	192	1,422	0.85	7
Kenya	167	167	1,430	0.98	10
Costa Rica	166	166	1,131	0.78	6

^aNote: Some running capacity figures were not available, and thus were assumed equal to the installed capacity

in geothermal water, can be extracted and used for carbonated beverages or to enhance growth in greenhouses. The typical equipment for a direct-use system is illustrated in Fig. 11, and includes downhole and

circulation pumps, heat exchangers (normally the plate type), transmission and distribution lines (normally insulated pipes), heat extraction equipment, peaking or back-up plants (usually fossil fuel fired)

Geothermal Energy Utilization. Table 5 National and regional geothermal power contributions

Country or region	% of national or regional capacity (MWe)	% of national or regional energy (GWh/year)
Lihir Island, Papua New Guinea	75	n/a
Tibet	30	30
San Miguel Island, Azores	25	n/a
Tuscany, Italy	25	25
Iceland	22	27
El Salvador	15	26
Kenya	12	17
Philippines	12	17
Nicaragua	11	10
Guadeloupe (Caribbean)	9	9
Costa Rica	8	12
New Zealand	6	10

to reduce the use of geothermal fluids and reduce the number of wells required, and fluid disposal systems (injection wells). Geothermal energy can usually meet 80 to 90% of the annual heating or cooling demand, yet only sized for 50% of the peak load. Geothermal heat pumps include both open (using groundwater or lake water) and closed-loop (either in horizontal or vertical configuration) systems as illustrated in Fig. 12.

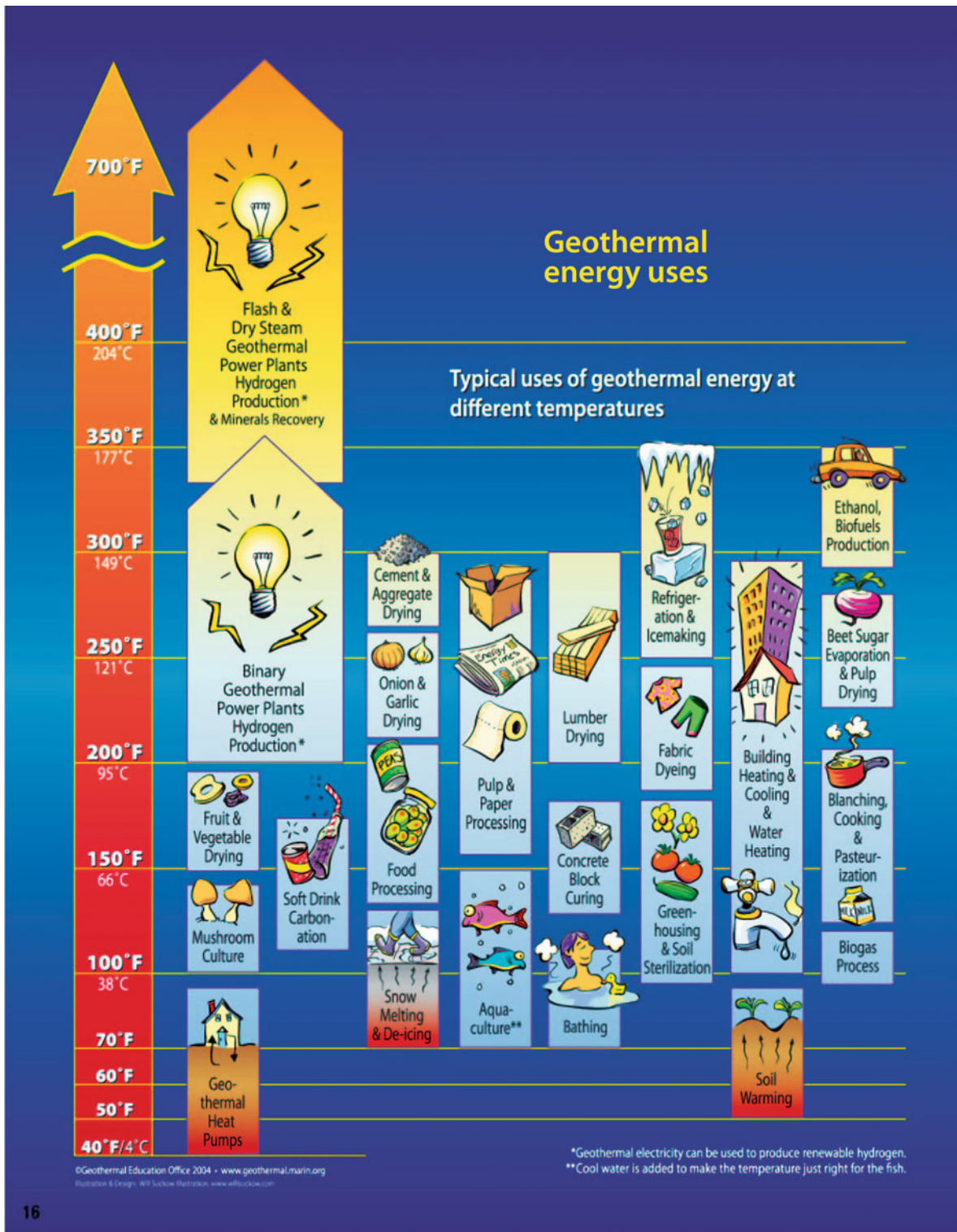
The world direct utilization of geothermal energy is difficult to determine; as, there are many diverse uses of the energy and these are sometimes small and located in remote areas. Finding someone, or even a group of people in a country who are knowledgeable on all the direct uses is difficult. In addition, even if the use can be determined, the flow rates and temperatures are usually not known or reported; thus, the capacity and energy use can only be estimated. This is especially true of geothermal waters used for swimming pools, bathing, and balneology.

One of the significant changes for WGC2010 was the increase in the number of countries reporting use. Six countries were added to the list in the current report as compared to 2005. In addition, the author is aware of three countries (Malaysia, Mozambique, and Zambia) that have geothermal direct uses, but did not provide a report for WGC2010. Thus, there are at least 81 countries with some form of direct utilization of geothermal energy.

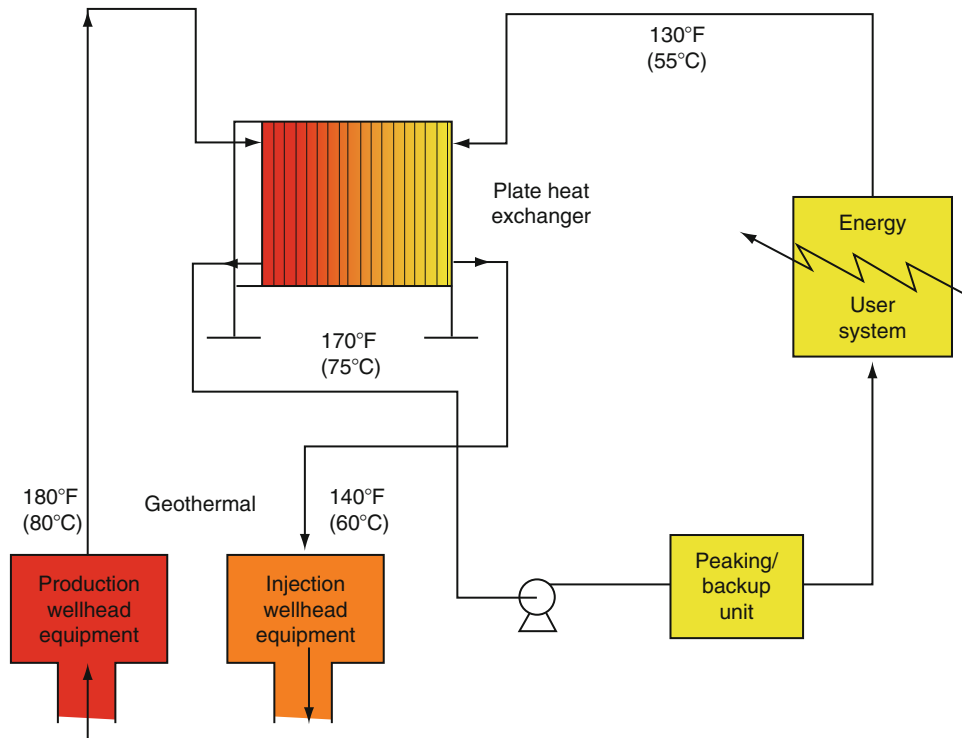
Another significant change from 2005 is the large increase in geothermal (ground-source) heat pump installations. They increased by 229% (18% annual growth) in capacity and 245% (20% annual growth) in annual energy produced over the 5-year period to the year 2010. At present (2010), they are the largest portion of the installed capacity (69.7%) and 49.0% of the annual energy use. The equivalent number of 12-kWt units installed (the average size) is approximately 3,000,000 in 43 countries, mostly in the United States, Canada, China, and Europe; however, the data are incomplete. The equivalent number of full-load heating operating hours per year varies from 2,000 in the United States, to over 6,000 in Sweden and Finland, with a worldwide average of 2,200 full-load h/year [10].

A summary of direct-use installed capacity and annual energy use are as follows (excluding geothermal heat pumps at 69.7% and 49.0% respectively of the total); bathing/swimming/spas 43.6% and 48.8%, space heating (including district heating) 35.1% and 28.2%; greenhouse heating 10.1% and 10.4%; aquaculture 4.3% and 5.2%; industrial 3.5% and 5.3%; agricultural drying 0.8% and 0.7%; cooling and snow melting 2.4% and 1.0%; and others 0.2% and 0.4%. District heating is approximately 85% of the space-heating use [10].

In terms of the contribution of geothermal direct use to the national energy budget, two countries stand out: Iceland and Turkey. In Iceland, it provides 89% of the country's space-heating needs, which is important since heating is required almost all year and saves about \$100 million in imported oil. Turkey has increased their installed capacity over the past 5 years from 1,495 to 2,084 MWt, most for district heating systems. A summary of some of the significant geothermal direct-use contributions to various countries is shown in Table 6, and the top direct-use countries are listed in Table 7 [10].

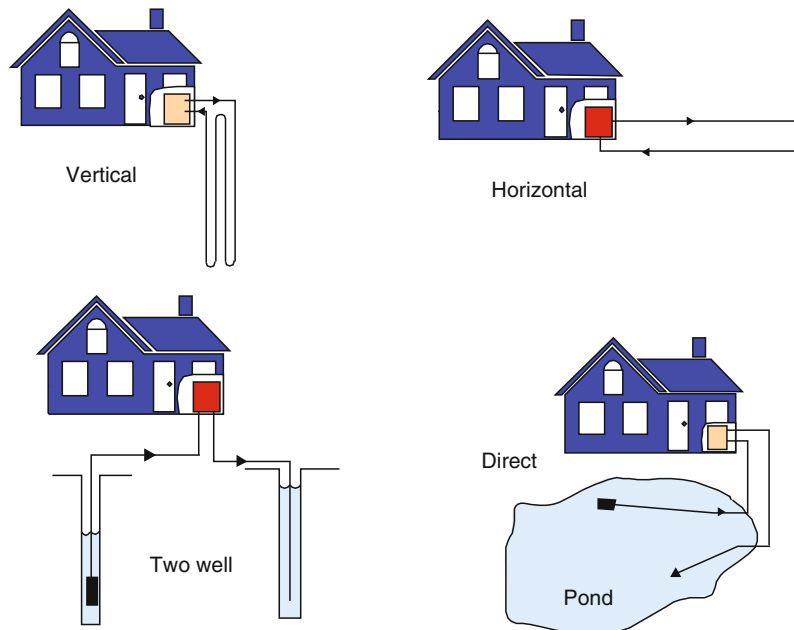


Geothermal Energy Utilization. Figure 10
Geothermal energy uses (Courtesy of the Geothermal Education Office)



Geothermal Energy Utilization. Figure 11

Typical direct-use geothermal heating system configuration



Geothermal Energy Utilization. Figure 12

Examples of common geothermal heat pump installations

Geothermal Energy Utilization. Table 6 National geothermal direct-use contributions

Iceland: provides 89% of country's space-heating needs
Turkey: space heating has increased 40% in the past 5 years, supplying 201,000 equivalent residences and 30% of the country will be heated with geothermal in the future
Tunisia: greenhouse heating has increased from 100 to 194 ha over the past 5 years
Japan: over 2,000 hot spring resorts (onsens), over 5,000 public bath houses, and over 15,000 hotels visited by 15 million quests per years, use natural hot springs
Switzerland: has installed 60,000 geothermal heat pumps = 1/km ² , and 2,000 km of boreholes were drilled in 2009. Drain water from tunnel are used to heat nearby villages and they have also developed several geothermal projects to melt snow and ice on roads
United States: has installed 1,000,000 geothermal heat pump units, mainly in the Midwestern and eastern states, with a 12.5% annual growth. Installation of these units is around 100,000–120,000/year

Environmental Considerations

Geothermal energy is considered a renewable and “green” energy resource; however, there are several environmental impacts that must be considered and are usually mitigated. These are emission of harmful gases, noise pollution, water use and quality, land use, and impact on natural phenomena, wildlife, and vegetation [13].

Emissions: These are usually associated with steam power plant cooling towers that produce water vapor emission (steam), not smoke. The potential gases that can be released, depending upon the reservoir type are carbon dioxide, sulfur dioxide, nitrous oxides, hydrogen sulfide, along with particulate matter. A coal-fired power plant produces the following kilograms of emissions per MWh as compared to a geothermal power plant: 994 vs. up to 40 for carbon dioxide, 4.71 vs. up to 0.16 for sulfur dioxide, 1.95 vs. 0 for nitrogen oxides, 0 vs. 0.08 for hydrogen sulfide (H₂S), and 1.01 vs. 0 for particulate matter. Hydrogen sulfide is routinely treated at geothermal power plants, and converted to elemental sulfur. In comparison, oil-fired power plants produce 814 kg and natural gas fired plants 550 kg of

Geothermal Energy Utilization. Table 7 Top Direct-Use Countries

Country	GWh/year	MWt	Main applications
China	20,932	8,898	Bathing/district heating
United States	15,710	12,611	GHP
Sweden	12,585	4,460	GHP
Turkey	10,247	2,084	District heating
Japan	7,139	2,100	Bathing (onsens)
Iceland	6,768	1,826	District heating
France	3,592	1,345	District heating
Germany	3,546	2,485	Bathing/district heating
Netherlands	2,972	1,410	GHP
Italy	2,762	867	Spas/space heating
Hungary	2,713	655	Spas/greenhouses
New Zealand	2,654	393	Industrial uses
Canada	2,465	1,126	GHP
Switzerland	2,143	1,061	GHP

H₂S/MWh. Binary power plants and direct-use projects normally do not produce any pollutants, as the water is injected back into the ground after use without exposing it to the atmosphere.

Noise: The majority of the noise produced at a power plant or direct-use site is during the well-drilling operation, which can shut down at night. The noise from a power plant is not considered an issue of concern, as it is extremely low, unless you are next to or inside the plant. Most of the noise comes from cooling fans and the rotating turbines.

Water use: Geothermal flash steam plants use about 20 l of fresh water/MWh, while binary air-cooled plants use no fresh water, as compared to a coal plant that uses 1,370 l/MWh. Oil plant use is about 15% less and nuclear about 25% more than the coal plant (www.cleanenergy.org). The only change in the fluid during use is to cool it, and usually the fluid is

Geothermal Energy Utilization. Table 8 Energy and greenhouse gas savings from geothermal energy production (electric at 35% efficiency and direct use at 70% efficiency)

	Fuel oil (10 ⁶)		Carbon (10 ⁶ t)			CO ₂ (10 ⁶ t)			SO _x (10 ⁶ t)			NO _x (10 ³ t)		
	Barrels	Tonnes	NG	Oil	Coal	NG	Oil	Coal	NG	Oil	Coal	NG	Oil	Coal
Electric	114	17	6	15	17	31	49	58	0	0.3	0.4	3.4	10.1	10.1
Direct use	154	23	9	23	27	46	74	88	0	0.5	0.5	4.5	13.6	13.6
Total	268	40	15	38	44	77	123	146	0	0.8	0.9	8.9	23.7	23.7

returned to the same aquifer so it does not mix with the shallow groundwater. At The Geysers facility in northern California, 42 million liters of treated wastewater from Santa Rosa are pumped daily for injection into the geothermal reservoir, reducing surface water pollution in the community and increasing the production of the geothermal field. A similar project supplies waste water (29 million liters daily) from the Clear Lake area on the northeast side of the The Geysers. These projects have increased the capacity of the field by about 200 MWe.

Land use: Geothermal power plants are designed to “blend-in” with the surrounding landscape, and can be located near recreational areas with minimum land and visual impacts. They generally consist of small modular plants under 100 MWe as compared to coal or nuclear plants of around 1,000 MWe. Typically, a geothermal facility uses 404 m² of land/GWh compared to a coal facility that uses 3,632 m²/GWh and a wind farm that uses 1,335 m²/GWh. Subsidence and induced seismicity are two land use issues that must be considered when withdrawing fluids from the ground. These are usually mitigated by injecting the spent fluid back into the same reservoir. There have been problems with subsidence at the Wairakei geothermal field in New Zealand; however, this has been checked by injection. Induced seismicity is also associated with EGS projects, producing earthquakes of less than 3.4 on the Richter scale. Neither of these potential problems is associated with direct-use projects, as the fluid use is small and well and pipelines are usually hidden. In addition, utilizing geothermal resources eliminates the mining, processing, and transporting required for electricity generation from fossil fuel and nuclear resources.

Impact on natural phenomena, wildlife, and vegetation: Plants are usually prevented from being located near geysers, fumaroles, and hot springs, as the extraction of fluids to run the turbines might impact these thermal manifestations. Most plants are located in areas with no nature surface discharges. If plants are located near these natural phenomena, the fluid extraction depth is planned from a different reservoir to prevent any impact. Designers and operators are especially sensitive about preserving manifestations considered sacred to indigenous people. Any site considered for a geothermal power plant, must be reviewed and considered for the impact on wildlife and vegetation, and if significant, provide a mitigation plan. Direct-use projects are usually small and thus have no significant impact on natural features.

In summary, the use of geothermal energy is reliable, providing base load power; is renewable; has minimum air emission and offsets the high air emissions of fossil fuel-fired plants; has minimum environmental impacts; is combustion free; and is a domestic fuel source.

Energy Savings

Using geothermal energy obviously replaces fossil fuel use and prevents the emission of greenhouse gases. If it is assumed that geothermal energy replaces electricity generation, the conversion efficiency is estimated at 0.35 (35%). These savings using geothermal energy at this efficiency level is summarized in Table 8 [14]. If the replacement energy for direct use is provided by burning the fuel directly, then about half this amount would be saved in heating

systems (35% vs. 70% efficiency), as used in Table 8. Savings in the cooling mode of geothermal heat pumps is also included in the figures in Table 8. The savings in fossil fuel oil is equivalent to about 3 days (1%) of the world's consumption.

It should be noted when considering these savings, that some geothermal plants do emit limited amounts of the various pollutants; however, these are reduced to near zero where gas injection is used and eliminated where binary power is installed for electric power generation. Since most direct-use projects use only hot water and the spent fluid injected, the above pollutants are essentially eliminated.

Future Directions

Geothermal growth and development of electricity generation has increased significantly over the past 40 years approaching 11% annually in the early part of this period, and dropping to 3% annually in the last 10 years due to the low price of competing fuels. Direct use has remained fairly steady over the 40-year period at 10% growth annually. The majority of the increase has been due to geothermal heat pumps. At the start of this 40-year period, only ten countries reported electrical production and/or direct utilization from geothermal energy. By the end of this period, 78 countries reported utilizing geothermal energy. This is almost an eightfold increase in participating countries. At least another ten countries are actively exploring for geothermal resources and should be on line by 2015.

Developments in the future will include greater emphases on combined heat and power plants, especially those using lower temperature fluids down to 100°C. This low-temperature cascaded use will improve the economics and efficiency of these systems, such as shown by those installed in Germany and Austria and at Chena Hot Springs, Alaska. Also, there is increased interest in agriculture crop drying and refrigeration in tropical climates to preserve products that might normally be wasted. Finally, the largest growth will include the installation and use of geothermal heat pumps, as they can be used anywhere in the world, as shown by the large developments in Switzerland, Sweden, Austria, Germany, China, Canada, and the United States.

Bibliography

Primary Literature

1. EPRI (Electric Power Research Institute) (1978) Geothermal energy prospects for the next 50 years. ER-611-SR, Special Report for the World Energy Conference 1978
2. Cataldi R, Hodgson S, Lund J (eds) (1999) Stories from a heated earth – our geothermal Heritage. Geothermal Resources Council, Davis, p 569
3. Lund JW (2006) Chena hot springs. Geo-Heat Center Quart Bull 27(3):2–4, Oregon Institute of Technology, Klamath Falls
4. Wright M (1998) Nature of geothermal resources. In: Lund JW (ed) Geothermal direct-use engineering and design guidebook. Geo-Heat Center, Klamath Falls, pp 27–69
5. White DE, Williams DL (eds) (1975) Assessment of geothermal resources of the United States – 1975. U.S. Geological Survey Circular 727, U.S. Government Printing Office, 155 p
6. Tenzer H (2001) Development of hot dry rock technology. Geo-Heat Center Quart Bull 22(4):14–22, Oregon Institute of Technology, Klamath Falls
7. Tester JW et al (2006) The future of geothermal energy – impacts of enhanced geothermal systems (EGS) on the United States in the 21st century. Massachusetts Institute of Technology, Cambridge, 384 p
8. Lund JW, Freeston DH (2001) World-wide direct uses of geothermal energy 2000. *Geothermics* 30(1):29–68, Elsevier, Oxford (updated and revised)
9. Lund JW, Freeston DH, Boyd TL (2005) Worldwide direct-uses of geothermal energy 2005. *Geothermics* 34(6):691–727, Elsevier, Amsterdam, The Netherlands
10. Lund JW, Freeston DH, Boyd TL (2010) Direct utilization of geothermal energy 2010 worldwide review. In: *Proceeding, World Geothermal Congress 2010, Bali, Indonesia*
11. Bertani R (2005) World geothermal general 2001–2005 – state of the art. *Geothermics* 34(6), Elsevier, Amsterdam, The Netherlands
12. Bertani R (2010) Geothermal power generation in the World, 2005–2010 update report. In: *Proceedings of the World Geothermal Congress 2010, Bali, Indonesia*
13. Kagel A, Bates D, Gawell K (2005) A guide to geothermal energy and the environment. Geothermal Energy Association, Washington, DC, 75 p
14. Goddard, W. B. and C. B. Goddard, 1990. "Energy Fuel Sources and Their Contribution to Recent Global Air Pollution Trends." Geothermal Resources Council *Transactions*, v. 14, Davis, CA, pp 643–649

Books and Reviews

- Armstead HCH (1983) Geothermal energy, 2nd edn. E. & F.N. Spon, London, 404 p
- Dickson MH, Fanelli M (2003) Geothermal energy utilization and technology. Earthscan, London, 205 p

- DiPippo R (2008) Geothermal power plants – principles, applications, case studies and environmental impact, 2nd edn. Elsevier, Amsterdam, 493 p
- Kavanaugh SP, Rafferty K (1997) Ground-source heat pumps – design of geothermal systems for commercial and institutional buildings. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc, Atlanta, 167 p
- Lund JW, Lienau PJ, Lunis BC (1998) Geothermal direct-use engineering and design guidebook, 3rd edn. Geo-Heat Center, Klamath Falls, 454 p

Websites

- European Geothermal Energy Council, Belgium: www.geothermie.de/egec_geothernet/menu/frameset.htm
- Geothermal Education Office, USA: <http://geothermal.marin.org>
- Geothermal Energy Association, USA: <http://www.geo-energy.org>
- Geo-Heat Center, USA: <http://geoheat.oit.edu>
- Geothermal Resources Council, USA: <http://www.geothermal.org>
- IEA (International Energy Agency) Heat Pump Center, Netherlands: www.heatpumpcentre.org
- International Geothermal Association: <http://www.geothermal-energy.org>
- International Ground Source Heat Pump Association, USA: <http://www.igshpa.okstate.edu>
- Stanford University Geothermal Program: <http://pangea.stanford.edu/ERE/research/geoth/>
- U.S. Department of Energy, Geothermal Technologies: www.eere.energy.gov/geothermal/
- World Geothermal Congress 2010, Indonesia: www.wgc2010.org

Geothermal Energy, Geology and Hydrology of

WILLIAM E. GLASSLEY

Energy Institute, University of California, Davis, CA, USA

Geologisk Institut, University of Aarhus, Aarhus, Denmark

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Heat Sources in the Earth
 Plate Tectonics as the Physical Framework
 Heat and Water in the Subsurface

Future Directions
 Summary
 Bibliography

Glossary

Core The central portion of the Earth that is composed of high density metallic, solid, and liquid components.

Crust The outer layer of the Earth composed of low to moderate density silicates and other minerals and within which the radioactive elements K, Rb, U, and Th are concentrated.

Direct use An application that uses the heat from a geothermal resource to accomplish heating, cooling and drying without converting thermal energy to another energy form.

Enhanced geothermal systems A deep geothermal system in which the porosity and permeability have been artificially enhanced through engineering methods to increase the mass flux of fluid that can be pumped through the reservoir.

Heat flow Strictly, the movement of thermal energy via diffusive conduction. Heat flow, as measured, is also a reflection of advective and convective transport.

Heat pump A device for transferring heat from one location to another.

Hydrology The scientific discipline that studies the flow of fluids in the crust.

Magma Molten rock that is one of the primary means for transferring heat to near-surface environments.

Mantle The interior portion of the Earth between the core and crust within which convective flow of material transfers heat to the crust.

Permeability The measurement or property of a medium that describes the ease with which a fluid will pass through the pores or fractures of the medium.

Plate tectonics The conceptual framework that provides a unifying principle describing the dynamic processes within the Earth.

Definition of the Subject

Geothermal energy is a ubiquitous renewable energy resource that is available virtually anywhere on the Earth. Surface manifestations of this energy resource

are, however, diverse and irregularly distributed. The most obvious and dramatic examples of geothermal energy are volcanoes. Less dramatic but equally unambiguous are geysers, hot springs, and warm pools, all of which are striking by their seemingly endless outflow of warm water from the subsurface. More subtle indications of geothermal energy are measurements in boreholes, mines, and wells that inevitably show that the deeper one goes below the surface, the warmer is the rock. All of these examples unambiguously document that heat is present in the subsurface, and it is this energy resource that geothermal applications utilize.

Access to geothermal resources varies from place to place, reflecting a complex interplay of geological and hydrological processes that have developed over millions of years. As a result, the types of geothermal applications that can be developed also vary from place to place. If high temperature (greater than $\sim 150^{\circ}\text{C}$) water can be accessed at depths of a few kilometers, the potential exists for installing a geothermal power plant. Lower temperature waters can be utilized for a broad range of so-called direct-use applications, whereby the thermal energy of the fluid is used for such things as drying timber, drying fruits and vegetables, curing concrete blocks, processing food, or heating buildings. And, virtually anywhere on the planet, at depths of a few meters to a few tens of meters, the constant flow of heat from the Earth's interior provides consistent conditions suitable for the installation of ground source heat pumps for heating and cooling. Evaluating the characteristics and magnitude of a geothermal resource requires unifying information, models, and concepts from a range of disciplines that focus on elucidating the properties of geological systems.

Introduction

Natural hot springs, volcanoes, and geysers are obvious indications that the interior of the Earth is hot. Catastrophic eruptions such as that of Mt. Vesuvius in AD 79 that destroyed the city of Pompeii and of Krakatoa in 1883 that affected weather patterns globally provide compelling evidence that the

magnitude of the heat energy is huge. But, the irregular distribution of volcanoes and hot springs over the Earth's surface seems, at first glance, to be enigmatic. If the interior of the Earth is hot, why are some manifestations of that heat restricted to certain regions? What controls the distribution pattern? How large is the resource?

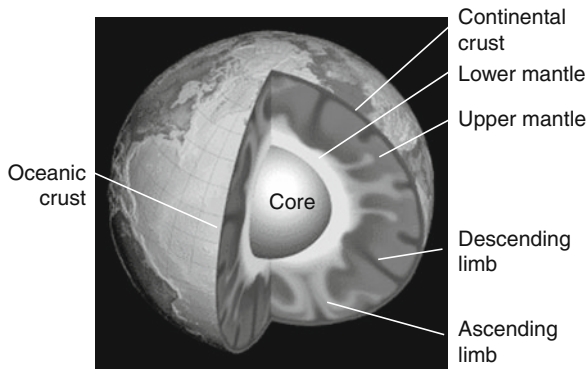
Answers to these questions derive from the evolution of geological processes.

Heat Sources in the Earth

Human awareness of geothermal energy dates back thousands of years [1] although the first uses of geothermal waters remain unknown. However, it was not until the Industrial and Scientific Revolutions in the 1700s and 1800s that investigations of the interior of the Earth began. Mining activities made it apparent that as one went deeper in the Earth temperatures increased [2, 3]. Why that should be so, and how hot the interior of the Earth was remained unknown until the advent of several scientific disciplines that, together, provided an answer to these questions.

The discovery of radioactivity in the late 1890s and early 1900s provided a solution to at least part of the problem. It was eventually recognized that radioactive decay generates heat. The Earth contains a number of radioactive elements, among them potassium (K), rubidium (Rb), uranium (U), and thorium (Th), all of which release heat when radioactive decay occurs. However, it was apparent that these elements could not account for the presence of a hot interior Earth in which there existed a liquid core, a fact that was generally accepted by the end of the 1920s [4]. These elements are concentrated in the crust of the Earth [5], and are of very minor importance in the mantle and core, and therefore they cannot be the exclusive source for the heat that is observed at the surface. This conundrum was resolved when the early accretionary history of the Earth became better understood.

The Earth accreted from the solar nebula about 4.56 billion years ago [6, 7]. The materials that formed the Earth included rocky and metallic bodies as well as icy material from comets. The kinetic energy



Geothermal Energy, Geology and Hydrology of.

Figure 1

A cross section of the Earth showing the main structural divisions. The core, lower mantle, upper mantle, and crust are indicated. Also shown are ascending and descending limbs of convection cells (Source: US Geological Survey, <http://geomag.usgs.gov/about.php>)

carried by these bodies when they impacted the forming planet was sufficient to heat the Earth substantially. At the same time, during the early life of the solar nebula an abundance of short-lived isotopes were also accumulating in the Earth. Particularly important were specific isotopes of aluminum, hafnium, and manganese (^{26}Al , ^{182}Hf , and ^{53}Mn , respectively). The combination of these heat-generating mechanisms ultimately resulted in the partial melting of the planet. Over a period of about 30 million years [8–13] metallic iron and related compounds became liquid and, because of their high density, settled to the core of the Earth while the remaining silicates stratified into layers (Fig. 1) of different densities. About 40% of the heat energy that is available and used in geothermal applications comes from this early period of differentiation of the Earth [14]. The remaining 60% comes from the decay of the longer-lived isotopes of K, U, Rb, and Th that have concentrated in the crust [15].

Plate Tectonics as the Physical Framework

The materials that make up the Earth are relatively poor thermal conductors. As a result, the heat deep in the Earth is conducted to the surface relatively

slowly. The average heat flow at the surface of the Earth is 87 milli-Joules/m²/s. Since a Watt (W) is a Joule per second (J/s), this heat flow is equivalent to 87 mW/m². Deeper in the Earth, near the core–mantle boundary, temperatures are in excess of 3,600°C [16]. The high temperatures in the low thermal conductivity materials that compose the mantle inevitably result in a situation where the lower mantle heats sufficiently to become less dense than the immediately overlying, cooler mantle. This is a gravitationally unstable configuration, and leads to upward flow of the hotter, less dense material. As a result, conditions favorable for development of a convection system become established [17]. The consequence of this condition, over time, is that hot mantle material begins to flow upward toward the Earth's surface [18]. When this ascending mantle material approaches the surface it spreads laterally, eventually descending to complete the pattern characteristic of convection. It is through this mechanism of convection that plate tectonics is driven, providing the resource for geothermal energy.

Plate tectonics describes the features and properties of the global convection system [19, 20]. The surface of the Earth is composed of seven major plates and approximately an equal number of smaller plates. The interior regions of the plates are relatively inactive, forming stable geological environments in which there is little seismic, tectonic, or volcanic activity. The edges of the plates, however, are the regions where seismic and tectonic activity are concentrated. It is mainly in these plate-edge settings where readily accessible, geothermally interesting resources occur. There are, in addition, several other geothermally important geological settings, most notably hot spots, that figure prominently in geothermal efforts. These tectonic environments localize heat in specific ways, providing an explanation for the observation that geothermal regions appear to be constrained to specific regions and zones around the world. Each of these settings is discussed in the following sections. The exceptions to this generalization are “geopressedured” resources and “Enhanced Geothermal Systems” (EGS) which are briefly discussed at the end of this chapter and are presented in detail in “► Engineered Geothermal Systems, Development and Sustainability of.”

Spreading Centers

When upward ascending limbs of convection cells approach the surface of the Earth, the hot material they are carrying begins to melt as the pressure drops. The melt aggregates into magma bodies that buoyantly rise. When the ascending limb is within about 100 km of the surface it begins to spread laterally, causing the plates on either side of this zone to move away from each other (Fig. 2). Magma in the ascending limb invades this zone, forming new crust at the edges of the diverging plates. Because of these magma bodies, spreading centers have the highest heat flow of any place on the Earth's surface. Theoretically, these zones could have heat flow values approaching $1,000 \text{ mW/m}^2$ [14].

Although of considerable theoretical interest from a geothermal energy perspective, spreading centers are of limited practical value since they are usually found in ocean basins at several kilometers depth and far from population centers. Exceptions to this are the Imperial Valley of California and the East African Rift. Both of these settings are geothermally active. Development of geothermal resources in Africa is underway. Geothermal power production in the Imperial Valley of California, which is currently

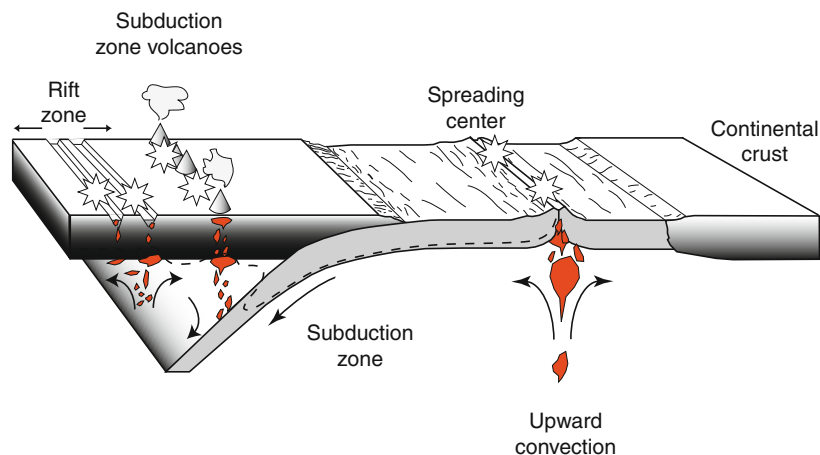
approximately 2,000 MW [21], provides a significant amount of the renewable energy generated by the state of California.

Subduction Zones

Once crust is formed, it is conveyed away from spreading centers and slowly cools. Since most newly formed crust forms in ocean basins, it interacts with seawater as it migrates away from the spreading centers and ages. This interaction with the seawater results in the formation of mineral phases that commonly contain some water in their structure.

As required by the law of conservation of mass, convecting systems invariable have descending limbs that counter the mass flow from ascending limbs. These zones in plate tectonics are called subduction zones (Fig. 2). Subduction zones are regions where the cooled crust descends back into the mantle. Because they are colder, they have low heat flow where the initial descent into the mantle occurs. In addition, as shown in Fig. 2, isotherms within the mantle are depressed in the immediate vicinity of the descending slab of cooler material.

Nevertheless, the descending crust does heat up as it enters the mantle. As the crust heats, it



Geothermal Energy, Geology and Hydrology of. Figure 2

Schematic diagram of the main elements in plate tectonic. The stars indicate regions where geothermal resources are concentrated. The dashed line schematically indicates the general form of isotherms. Magma bodies are indicated by the red forms

eventually reaches temperatures sufficient to cause the breakdown of the hydrous minerals that formed when the crust was interacting with seawater. This breakdown of hydrous minerals liberates water. Because of its relatively low density, the released water migrates upward into the overlying warmer mantle where it causes a complex series of reactions, including partial melting. The melts that are generated ascend to the surface, where they form prominent volcanic chains. The so-called Ring of Fire that surrounds the Pacific Ocean formed precisely as a result of this process. This process is, in essence, a heat transfer mechanism whereby the rising magma brings heat to the near-surface environment from the deeper mantle.

The volcanic systems associated with subduction zones are, globally, the most common settings for geothermal power generation and direct use. Geothermal facilities that have utilized the thermal energy in these settings have been built in Chile, Central America, the Cascades in Northern California and Oregon, Japan, the Philippines, and the Mediterranean region, to name a few.

Commonly associated with the volcanic chain that forms above subduction zones are secondary rift systems, or back-arc basins (Fig. 2). These environments are extensional systems somewhat like spreading centers [22]. They appear to develop in response to complex flow dynamics in the mantle above the descending slab [23]. Because they involve the same type of ascending hot mantle flow as found in spreading centers, magma bodies form and ascend to the near surface. As a result, these settings can also be important high temperature zones within which geothermal resources concentrate. An example of this type of environment is the Taupo volcanic zone on the North Island of New Zealand, where occurred the first large-scale development of geothermal power in the world. Although geologically more complex, the Basin and Range Province of the Western United States is, in part, a manifestation of similar processes. Existing and/or planned geothermal facilities in eastern Arizona, California, Colorado, Idaho, Montana, Nevada, and New Mexico, to name a few, utilize geothermal resources in this type of setting.

Transform Faults

The third type of plate boundary is a feature called a transform fault. Transform faults are places where plates move past each other horizontally, forming fault zones in which the rock has been crushed and fragmented. Although transform faults are not intrinsically associated with magma bodies, local geological conditions can result in high heat flow and the development of a geothermal resource. This can occur because such zones provide easy flow paths for warm or hot waters at depth to ascend to shallower levels. The San Andreas fault in California is one example of a transform fault along which warm and hot springs are common and for which there is chemical and isotopic evidence of flow from great depth [24].

Hot Spots

Hot spots are surface manifestations of a persistent heat source deep in the mantle. They are characterized by long-lived volcanic activity and by the fact that they exhibit no specific relationship with any type of plate boundary. Hot spots are found relatively commonly in the interiors of plates. Their cause remains largely unknown. Iceland and the Hawaiian Island–Emperor Seamount chain are two examples of hot spots. In the case of Iceland, the heat and magma source for the hotspot coincides with the spreading center in the Atlantic Ocean. It has persisted at that site for more than 10 million years. The Hawaiian–Emperor chain, by contrast, has traveled over a hot spot that has persisted for more than 50 million years. The only active volcano in this chain is the island of Hawaii. Both of these locations have geothermal facilities, Iceland in particular being a spectacular example of the broad use of geothermal energy resources for both space and district heating and power generation.

Heat and Water in the Subsurface

The above discussion provides a broad conceptual framework for understanding how geothermal heat becomes concentrated in certain regions. Generally,

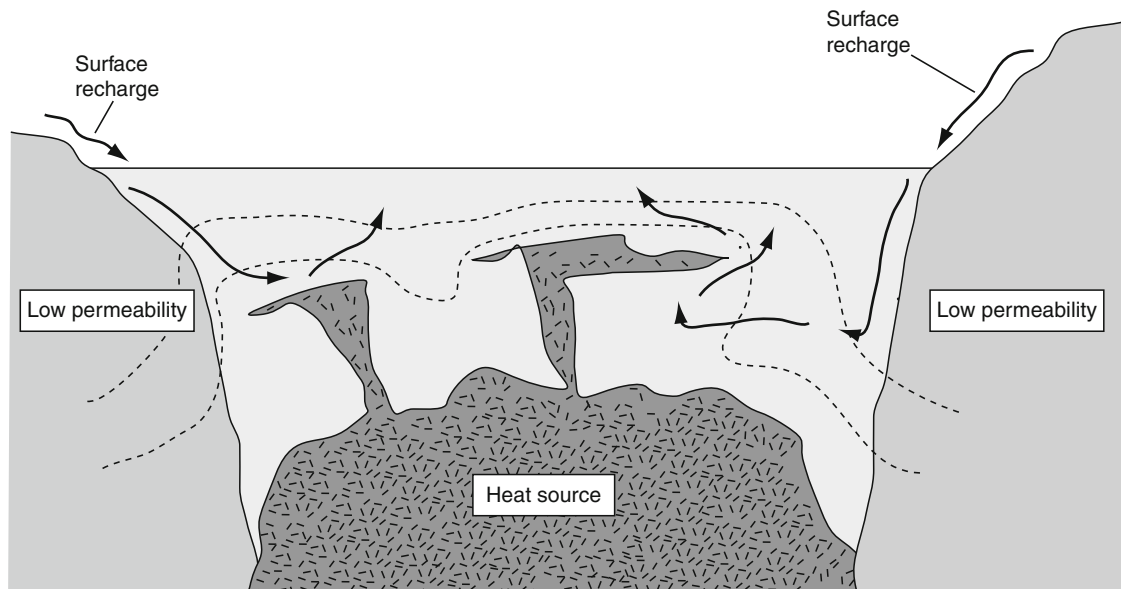
the responsible processes transport heat to the near surface either through the intrusion of magma or the flow of deep circulating water. The following discussion considers in more detail the specific geological and hydrological conditions that make a heat resource useful for geothermal applications.

Heat Sources

Geothermal power generation uses the heat energy of moderate to high temperature ($>140^{\circ}\text{C}$) geothermal fluids to power turbines that drive electrical generators (see “► [Geothermal Power Conversion Technology](#)”). Direct-use applications, such as aquaculture, food drying, district heating systems, and greenhouses, among others, rely on lower temperature (less than about 150°C) geothermal fluids to directly heat an environment for specific purposes (see entries “► [Nuclear Transfer to Produce Transgenic Mammals](#)” and

“► [Disease-Resistant Transgenic Animals](#)”). Regardless of the technology employed, there are several basic geological and hydrological conditions that determine whether a resource is suitable for its intended use. The key requirements are an adequate resource that can provide the requisite thermal energy for the specific application and whether there is an adequate fluid flow to transfer the thermal energy to power the engineered system.

Cooling magma bodies, or their hot solidified counterparts, are the heat source for many geothermal applications. These bodies can be of many forms and sizes. Shown in [Fig. 3](#) is an example of the kind of irregularity they may have. The main magma body (pluton) can be a few hundred meters to kilometers in size, with offshoots (dikes and sills) a few meters to many tens of meters in size. The magma, when it is liquid, will have temperatures between about 700°C and $1,100^{\circ}\text{C}$, depending on its composition.

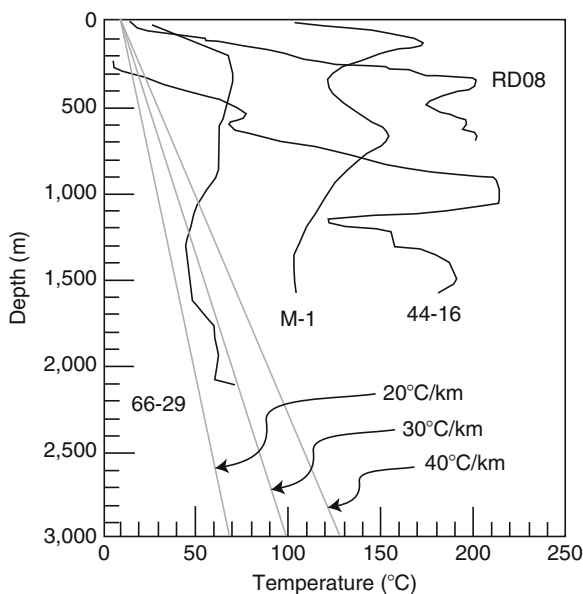


Geothermal Energy, Geology and Hydrology of. Figure 3

Schematic diagram of a high temperature geothermal system. The *arrows* indicate the flow path of water in the subsurface. Hypothetical 250°C and 300°C isotherms are indicated by the *dashed lines*. The *dark gray* body labeled "Heat source" represents a cooling igneous body that once was a magma chamber. The *light gray* pattern indicates porous and permeable rocks into which the magma intruded. The *medium gray* rocks labeled "Low permeability" indicate highland regions that provide recharge of water to the subsurface

The cooling rate for shallow plutons is tens to hundreds of degrees per million years; hence, the lifetime of useful heat output can be quite long. This allows geothermal development of intrusive bodies that are several to ten million years old, depending on the local conditions.

The distribution of heat around these bodies can result in complex isotherm patterns. Shown in Fig. 3 are hypothetical 300°C and 250°C isotherms that might be expected solely from slow cooling of the heat source. The form of the isotherms is influenced by the geometry of the igneous body, as well as the local geology. As will be discussed below, another important factor that influences isotherm form is the extent to which fluid flow transfers heat away from the region. Regardless of the underlying mechanism, the isotherms in such a setting are likely to have an irregular form. It is for this reason that temperature gradient holes are used in exploration efforts to determine the real vs hypothetical subsurface temperature distribution [25]. The importance of acquiring such information is shown in Fig. 4, which



Geothermal Energy, Geology and Hydrology of.

Figure 4

Measured temperature profiles in bore holes at Long Valley caldera, California (Data from [26])

depicts measured temperatures in the subsurface in Long Valley caldera in the eastern Sierra Nevada mountains [26]. The Long Valley caldera is a region of geothermal activity and is the site of a 37 MW power generation facility. Linear geothermal gradients are shown for comparison. Note the pronounced departure from linearity. The prominent temperature spikes, as well as the complex variations in temperature over distances of several hundred meters reflect the effects of fluid flow.

The other main source of useful geothermal energy is deep circulation of water in natural aquifer systems. In this instance, meteoric water flows into the subsurface, often along faults or other flow paths. The situation is exactly analogous to the flow field depicted on the right hand side of Fig. 3, except in this case there is no specific heat source present. Instead, the descending water flows to depths of several kilometers where the normal geothermal gradient (10–30°C/km) results in the circulating fluid coming in to contact with rocks in the temperature range of 50–100°C. Deep fault zones that intersect these naturally circulating fluids can allow rapid ascent of the fluid from depth, resulting in the development of warm or hot springs. Although such fluids do not possess sufficient energy to support power generation, they do have sufficient thermal energy for successful development of direct-use applications.

Whether or not a high temperature zone will be useful for geothermal applications depends on whether the heat it possesses can be brought to the surface in sufficient quantity. The fundamental requirement for achieving this is the ability to circulate fluid in sufficient volume to bring the heat to the surface at a rate that matches the demand of the application. The material property that determines whether this criteria can be met is the permeability.

Subsurface Fluid Flow: Porosity and Permeability

Rocks are generally classified as igneous, metamorphic, or sedimentary, depending upon the principal process that controlled their formation. Igneous rocks crystallized from magmas and usually have a very low percentage (usually much less than 10%)

of their volume occupied by pore space (*porosity*). Sedimentary rocks form as the products of erosion and deposition, often in response to the effects of water movement or settling in sedimentary basins. Such rocks have a wide range of porosities, but can be quite porous, with values easily reaching 40% or more. Metamorphic rocks form as a result of changes in temperature and pressure due to burial or other physical effects. As the physical conditions evolve a rock will recrystallize, changing its mineralogy and internal structure. The porosity of metamorphic rocks usually falls somewhere between that of igneous and sedimentary rocks.

All rocks have two related but independent physical properties. One of these properties is porosity (as described above), the other is permeability. Regardless of how solid a rock appears, there will always be some amount of space between mineral grains and/or some cracks and fractures. If the pore space occurs primarily as voids between grains, the porosity is classified as *matrix-dominated porosity*. If most of the void space occurs as fractures, it is termed *fracture-dominated porosity*. The open space between mineral grains can be as small as a micron (i.e., 1 millionth of a meter) or as large as a significant fraction of a centimeter. Fractures can have similar dimensions. Measured values of rock porosity vary from a small fraction of a percent in unfractured crystalline rocks such as granites or some metamorphic rocks, to greater than 40% in some sedimentary sandstones. The porosity determines the instantaneous volume of fluid a rock can possess.

The porosity has an important effect on the thermal and mechanical properties of a rock in the subsurface. The thermal properties of water and minerals are significantly different. At 25°C, for example, the amount of energy it takes to raise the temperature of 1 kg of water 1°C is 4,180 J. On the other hand, it only takes 660 J of energy to raise the temperature of 1 kg of potassium feldspar, a common mineral in granite, 1°C. In other words, the heat capacity of water is quite high compared to the heat capacity of many minerals. Thus, at a given temperature, assuming the porosity is 100% filled by liquid water (i.e., the rock is *saturated*), the heat content of a given volume of potassium feldspar with 1% porosity will be about half that for the case in which

the porosity is 20%. Similar results are obtained for most other minerals. Heat capacity varies with mineral, and is a function of temperature. Hence, this specific result *qualitatively* illustrates the relationship between heat content, porosity, and water content (or saturation) for rock systems, but does not *quantitatively* represent the behavior of all rocks under these conditions. When assessing the available thermal energy in a potential geothermal resource, it is important to establish the porosity and degree of saturation of the rock composing the geothermal reservoir. A detailed discussion of resource assessment is provided in later entries in this book.

The ability of fluid to flow through a geothermal reservoir determines the rate at which heat can be extracted. The measure of the ease with which fluid flows through rock is the *permeability*. For an unfractured, porous, saturated rock, if the pores are not interconnected the fluid cannot flow through the rock and the permeability will necessarily be zero. In other words, regardless of the porosity, there would be no fluid flow in this case. For porous rock with interconnected porosity, the ability of fluid to flow will depend on the size of the connections between the pores, the complexity of the flow path (also called the *tortuosity*), the pressure gradient across the flow path, the shape of the pores, and certain physical properties of the fluid. For fractured rocks [27], the important characteristics affecting the permeability include the effective aperture of the fracture, its roughness, the number of fractures per rock volume, and their orientation, as well as the fluid properties.

The concept of permeability was formalized by Henry Darcy in the 1800s. He developed the relationship

$$q = -(\kappa/\mu) \cdot A \cdot \nabla(P)$$

where q is the flux ($\text{m}^3/\text{m}^2/\text{s}$), κ is the permeability (in units of area, m^2), A is the cross-sectional area (m^2), μ is the dynamic viscosity ($\text{kg}/(\text{m}\cdot\text{s})$), and $\nabla(P)$ is the gradient in pressure. This relationship strictly applies only to conditions of slow flow in porous media for a single phase [28, 29]. It is often used in more complex situations, but its limitations need to be understood. It is especially important to recognize this limitation in geothermal systems where high flow rates

are often encountered. Note that a standard measure of permeability is the darcy, the conversion for which is $9.869 \text{ e}^{-13} \text{ m}^2/\text{darcy}$.

The relationship between porosity and permeability has been formalized in several ways [30], the most useful of which is the Kozeny–Carmen equation [31–33],

$$\kappa = [n^3/(1-n)^2]/(5 \cdot S_A)^2$$

In this relationship, n is the porosity and S_A is the surface area of the pore spaces per unit volume of rock. This relationship can also be written as

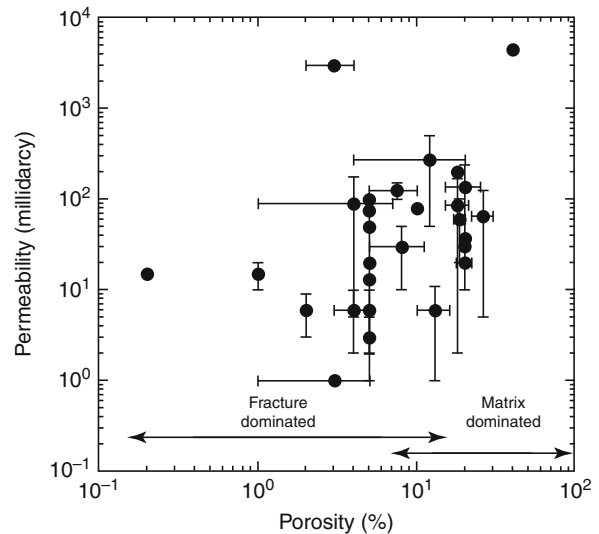
$$\kappa = c_0 \cdot T[n^3/(1-n)^2]/(S_A)^2$$

where T is the tortuosity and c_0 is a constant. The tortuosity is the ratio between the straight path between two points and the actual path length a particle would follow in the flow field. Commonly $c_0 T$ is treated as equivalent to 0.2, making it identical to the Kozeny–Carmen equation.

The ability to obtain sufficient energy for use in geothermal applications usually requires flow rates on the order of several cubic meters per second. Hence, it is important that drilling programs target regions in the subsurface with at least moderate permeability.

Permeability in Geothermal Systems

The physical properties of rocks in real geothermal systems are highly variable. Some geothermal systems extract energy from highly porous sandstones and other sedimentary rocks, while others utilize geothermal resources that occur in fractured crystalline rocks. Björnsson and Bodvarsson [34] compiled porosity and permeability data from operating geothermal power plant locations in various geological settings and documented the high degree of variability such systems possess (Fig. 5). Despite the variability, it is apparent that systems in which the porosity is primarily in the form of matrix pores rather than fractures, a much higher porosity is required to achieve sufficient flow to support power generation. Fracture-dominated systems, on the other hand, can have very low overall porosity and permeability yet support sufficient flow to generate power. This observation suggests that in a fractured rock mass,



Geothermal Energy, Geology and Hydrology of.

Figure 5

Observed porosity vs permeability in geothermal systems used for power generation (Data from [34])

fluid flow may be concentrated in a very small proportion of the overall fracture population [27] and yet still be adequate for supporting power generation. This conclusion is supported by modeling that has recently been reported [35].

Using a fractal model for fracture properties, along with data from well-characterized geothermal systems, Williams [35] demonstrated that the bulk of fluid flow in these geothermal systems occurs in a small fraction of the total porosity. This implies that a few fractures out of a population of many fractures carry most of the fluid flow. These results emphasize the importance of sufficiently characterizing rock properties at a site to establish the likely range of permeability properties. Once established, this information can be used to determine drilling targets and likely production levels.

Future Directions

Basic Geological Principles for Enhanced Geothermal Systems (EGS)

Implicit in the discussion of the driving forces for plate tectonics is the fact that heat is present everywhere in

the subsurface. Geothermal gradients of a few degrees to several tens of degrees per kilometer lead to the conclusion that temperatures greater than 300°C, which are sufficient to generate power, can be obtained at depths between 5 and 15 km virtually anywhere in the world. Pursuit of this resource has a long history [36–38]. Indeed, a recent study [39] concluded that in the United States alone, the amount of thermal energy that could be accessed at depths of less than 10 km is in excess of 13 million exajoules (1 exajoule = 10^{18} J). If only 1.5% of that energy could be accessed it would supply more than 2,000 times the annual electrical power generation needs of the country. Similar conclusions apply for almost every country on the planet.

The challenges faced in accessing this energy are the depths to which drilling must routinely go to tap the resource, and the ability to bring the heat to the surface. The depth to be drilled is determined by the regional geological framework. The interiors of plates often are stable environments that have had little magmatic activity. In the absence of magmatic activity, normal geothermal gradients in the interiors of plates are relatively low, on the order of a few degrees per kilometer to about 20°/km. These conditions would require drilling to depths that approach the limits of current drilling technology. Nevertheless, all plates have large areas in which adequate temperatures can be accessed at depths less than 10 km, and represent potential drilling targets for EGS systems.

The other challenge faced by EGS development is the ability to circulate fluids to depths in sufficient volumes to extract useful quantities of heat. Deep boreholes tend to enter regions where the permeability is low and the volume of subsurface water is small. To overcome these problems, it is possible to enhance the permeability of the rock using standard techniques of hydrofracturing that have been practiced in the oil and gas industry for decades. Hydrofracturing allows development of sufficient fracture permeability to support fluid flow at the volumes required for power generation. For a detailed discussion of EGS technology, see “► [Engineered Geothermal Systems, Development and Sustainability of.](#)”

Basic Geological Principles for Non-power Producing Applications

Direct-use applications and ground source heat pumps require much lower temperatures than those needed for power generation [40]. As a result, they can usually rely on resources that are in the relatively shallow subsurface. Direct-use applications, such as aquaculture, spas, food processing, lumber drying, etc., often are located where warm or hot water occurs naturally in springs or the immediate subsurface. Such settings mainly rely on high permeability zones, such as local faults or porous and fractured rocks such as some volcanic deposits, for the fluid supply. As a result, most such applications are located in regions where recent volcanic activity has provided a thermal resource.

The use of ground source heat pumps for space heating and cooling [41], on the other hand, is not restricted by location. Ground source heat pumps utilize technology similar to that employed to cool the interior of refrigerators. Using technologically sophisticated but mechanically simple heat pumps, these systems can extract heat from the subsurface and transfer it to a building space (for space heating), or remove heat from a building space and deposit it in the subsurface (for cooling). Because of their high efficiencies, these heat pumps require a thermal resource of only 10–15°C (ca. 50–60°F). Such temperatures are readily accessed within 30–100 m in the subsurface, due to the normal geothermal gradient. These systems can be installed in virtually any geological setting. However, their long-term performance is influenced by the thermal properties of the rock types in a given area, as well as the presence or absence of subsurface water. If an active aquifer is located at the same site, the thermal stability of the system can be influenced by the rate of flow and recharge area of the aquifer. Shallower systems, such as trench installations, are also possible in some areas if suitable space is available and if appropriate soil depths occur.

Summary

Geothermal resources are a reflection of the underlying global and local geological and hydrological framework. The most thermally rich resources tend

to concentrate in environments that have abundant volcanic activity. These tend to be controlled by plate tectonic processes and are, specifically, spreading centers, volcanic chains associated with subduction zones and hot spots. The local geological characteristics that favor useful resources include relatively shallow depths to the resource, high permeability in the rocks surrounding the resource, and adequate fluids. These conditions apply for all applications except those utilizing ground source heat pumps. For these systems, virtually any setting is suitable, since the normal flow of heat from the Earth is adequate to assure a thermal resource of a 10–20°C within a few tens to a few hundreds of meters in the subsurface.

Bibliography

Primary Literature

- Cataldi R (1993) Review of historiographic aspects of geothermal energy in the Mediterranean and Mesoamerican areas prior to the modern age. *Geoth Heat Cent Bull* 15:13–16
- Buffon GL (1778) *Histoire naturelle, générale et particulière*. Imprimerie Royale, Paris
- Dickson MH, Fanelli M (2006) Geothermal background. In: Dickson MH, Fanelli M (eds) *Geothermal energy: utilization and technology*. Earthscan, London
- Brush SG (1979) Nineteenth-century debates about the inside of the earth: solid, liquid or gas? *Ann Sci* 36:225–254
- Van Schmus WR (1995) Natural radioactivity of the crust and mantle. In: Ahrens TJ (ed) *Global earth physics*. American Geophysical Union, Washington, DC, pp 283–291
- Göpel C, Manhés G, Allégre CJ (1994) U-Pb systematics of phosphates from equilibrated ordinary chondrites. *Earth Planet Sci Lett* 121:153–171
- Allégre CJ, Manhés G, Göpel C (1995) The age of the Earth. *Geochim Cosmochim Acta* 59:1445–1456
- Wetherill GW (1990) Formation of the Earth. *Annu Rev Earth Planet Sci* 18:205–256
- Canup RM, Agnor C (2001) In: Canup RM, Righter K (eds) *Origin of Earth and moon*. Cambridge University Press, Cambridge, pp 1839–1848
- Chambers JE (2001) Making more terrestrial planets. *Icarus* 152:205–224
- Kortenkamp SJ, Wetherill GW, Inaba S (2001) Runaway growth of planetary embryos facilitated by massive bodies in a protoplanetary disk. *Science* 293:1127–1129
- Kleine T, Münker C, Mezger K, Palme H (2002) Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf-W chronometry. *Nature* 418:952–955
- Yin Q, Jacobsen SB, Yamashita K, Blichert-Toft J, Te'louk P, Albarede F (2002) A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* 418:949–952
- Stein CA (1995) Heat flow in the Earth. In: Ahrens TJ (ed) *Global earth physics*. American Geophysical Union, Washington, DC, pp 144–158
- Shih K-G (1971) Temperature production in the continental crust due to radioactive heat production. *Pure Appl Geophys* 90:115–125
- Alfé D, Gillian MJ, Price GD (2007) Temperature and composition of the Earth's core. *Contemp Phys* 48:63–80
- Anderson DL (1989) *Theory of the Earth*. Blackwell, Boston
- Yamazaki D, Karato S-I (2001) Some mineral physics constraints on the rheology and geothermal structure of Earth's lower mantle. *Am Mineralog* 86:385–391
- Isacks B, Oliver J, Sykes LR (1968) Seismology and the new global tectonics. *J Geophys Res* 73:5855–5899
- Glen W (1982) *The road to Jaramillo: critical years of the revolution in earth science*. Stanford University Press, Stanford
- CGEC (California Geothermal Energy Collaborative) (2006) *California geothermal fields and existing power plants, Fact Sheet*
- Hart SR, Glassley WE, Karig DE (1972) Basalts and sea-floor spreading behind the Mariana island arc. *Earth Planet Sci Lett* 15:12–18
- Hirth G, Kohlstedt D (2003) Rheology of the upper mantle and the mantle wedge: a view from the experimentalists. *Geophys Monograph* 138:83–105, American Geophysical Union, Washington, DC
- Kennedy BM, Kharaka YK, Evans WC, Ellwood A, DePaolo DJ, Thordsen J, Ambats G, Mariner RH (1997) Mantle fluids in the San Andreas fault system, California. *Science* 278:1278–1281
- Lachenbruch AH (1968) Preliminary geothermal model of the Sierra Nevada. *J Geophys Res* 73:6977–6989
- Farrar CD, Sorey ML, Roeloffs E, Galloway DL, Howle JF, Jacobson R (2003) Inferences on the hydrothermal system beneath the resurgent dome in Long Valley Caldera, east-central California, USA, from recent pumping tests and geochemical sampling. *J Volcanol Geoth Res* 127:305–328
- Bear J (1993) Modeling flow and contaminant transport in fractured rocks. In: Bear J, Tsang C-F, de Marsily G (eds) *Flow and contaminant transport in fractured rock*. Academic, New York
- Bear J (1979) *Hydraulics of groundwater*. McGraw-Hill, New York
- Batu V (1998) *Aquifer hydraulics*. Wiley, New York

30. Lee CH, Farmer IW (1990) A simple method of estimating rock mass porosity and permeability. *Int J Min Geol Eng* 8:57–65
31. Kozeny J (1927) Über kapillare Leitung des Wassers im Boden. *Sitzungsber. Akad Wissenschaft Wien* 136: 271–306
32. Carman PC (1937) Fluid flow through a granular bed. *Trans Inst Chem Eng Lond* 15:150–156
33. Carman PC (1956) *Flow of gases through porous media*. Butterworths, London
34. Björnsson G, Bodvarsson G (1990) A survey of geothermal reservoir properties. *Geothermics* 19:17–27
35. Williams CF (2007) Updated methods for estimating recovery factors for geothermal resources. In: *Proceedings of the 32nd Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, California
36. Smith MC (1983) A history of hot dry rock geothermal energy systems. *J Volcanol Geoth Res* 15:1–20
37. Tester JW, Brown DW, Potter RM (1989) Hot dry rock geothermal energy – a new energy agenda for the 21st Century. Los Alamos National Laboratory report LA-11514-MS
38. Duchane D, Brown D (2002) Hot dry rock (HDR) geothermal energy research and development at Fenton Hill, New Mexico. *Geo-Heat Center Bull* 23(4):13–19
39. Tester JW, Anderson BJ, Batchelor AS, Blackwell DD, DiPippio R, Drake EM, Garnish J, Livesay B, Moore MC, Nichols K, Petty S, Toksoz MN, Veatch RW Jr (2006) *The future of geothermal energy*. MIT Press, Boston
40. Lund JW, Freeston DH, Boyd TL (2005) Direct application of geothermal energy. 2005 worldwide review. *Geothermics* 34:690–727
41. Ochsner K (2008) *Geothermal heat pumps*. Earthscan, London

Books and Reviews

- DiPippio R (2008) *Geothermal power plants*, 2nd edn. Elsevier, Amsterdam
- Duffield WA, Sass JH (2003) *Geothermal energy – clean power from the Earth's heat*. U.S. Geological Survey Circular 1249
- Glassley WE (2010) *Geothermal energy: renewable energy and the environment*. Taylor and Francis, Boca Raton
- Krauskopf KB, Bird DK (2003) *Introduction to geochemistry*, 3rd edn. McGraw-Hill, New York
- Reynolds JM (1997) *An introduction to applied and environmental geophysics*. Wiley, New York
- Ryback L, Muffler LJP (1979) *Geothermal systems: principles and case histories*. Wiley, New York
- Williams CF, Reed MJ, Mariner RH (2008) A review of methods by the U.S. Geological Survey in the assessment of identified geothermal resources. U.S. Geological Survey Open File Report 2008-1296

Geothermal Energy, Nature, Use, and Expectations

BARRY GOLDSTEIN¹, GERARDO HIRIART², JEFF TESTER³,
LUIS GUTIERREZ-NEGRIN⁴, RUGGERO BERTANI⁵,
CHRISTOPHER BROMLEY⁶, ERNST HUENGES⁷,
ARNI RAGNARSSON⁸, MIKE MONGILLO⁶, JOHN W. LUND⁹,
LADISLAUS RYBACH¹⁰, VLADIMIR ZUI¹¹,
HIROFUMI MURAOKA¹²

¹Pirsa Petroleum Group, Adelaide, SA, Australia

²Energías Alternas, Estudios y Proyectos, Mexico

³Energy Institute, Cornell University, Ithaca, NY, USA

⁴Mexican Geothermal Association, Mexico

⁵Enel, Rome, Italy

⁶GNS Science, Wairakei Research Centre, Taupo, New Zealand

⁷GFZ-Potsdam, Germany

⁸Iceland GeoSurvey, Reykjavík, Iceland

⁹Geo-Heat Center, Oregon Institute of Technology, Klamath Falls, OR, USA

¹⁰Geowatt AG, Switzerland

¹¹Belarusian Research Geological Prospecting Institute, Belarus

¹²National Institute of Advanced Industrial Science and Technology, Institute for Geo-Resources and Environment (GREEN), Tsukuba, Ibaraki, Japan

Article Outline

Glossary

Definition of Geothermal Energy

Introduction

Geothermal Resources, Reserves, and Supplies

Future Directions for Geothermal Energy Technologies

Expectations for Geothermal Energy Use

Key Points

Acknowledgments

Bibliography

Glossary

Base-load demand Continuous demand for electricity. Power generation plants with high-capacity factors combine as a practical source of continuous base-load supplies.

Capacity factor The energy generated in a span of time divided by the maximum energy that could have been generated at full (name plate) power of the plant during that period of time, most often expressed as a percentage of 1 year of plant operation. The maximum amount of power a plant can generate is its name plate capacity.

Conduction-dominated systems Earth systems of heat transfer in which heat flow is principally via the contact of rocks (and pore- and fracture-filling fluids and gasses in rocks) with a capacity to transfer thermal energy from higher to lower temperature conditions. Non-volcanic (amagmatic) geothermal systems tend to become conduction-dominated systems.

Convection-dominated systems Earth systems of heat transfer in which heat flow is principally via flow of gasses, fluids, and molten rock (magma) from higher to lower temperature conditions. Volcanic geothermal systems tend to become convection-dominated systems.

Dispatchable electricity Power generation systems that can quickly shift from nil to full generation capacity and balance electricity supply and demand within safe technical limits of transmission grids.

Engineered (or enhanced) geothermal systems (EGS) Geothermal reservoirs in which technologies enable economic utilization of low permeability conductive dry rocks or low productivity convective water-bearing systems by creating fluid connectivity through hydraulic, thermal, or chemical stimulation methods or advanced well configurations. EGS also refer to activities to increase the permeability in a targeted subsurface volume via injecting and withdrawing fluids into and from the rock formations that are intended to increase the ability to extract energy from a subsurface heat source.

Geothermal energy Accessible thermal energy stored in the Earth's interior, in rock, gasses, and fluids usable for the generation of electricity and to supply heat for direct use. Continuous radiation from the natural decay of elements and residual energy from the earth's formation are the main sources of geothermal energy.

Ground source heat pumps Equipment that circulates fluids or gasses from lower to higher temperature conditions, or the reverse to heat or cool

buildings or industrial processes. Ground source heat pumps (GSHPs) are most commonly used to heat in winter and cool in summer.

Hot sedimentary aquifers Any geologic reservoir that has a capacity to flow fluids at a rate and a temperature sufficient to meet a market for power generation and the direct use of thermal energy. The most accessible and the most prospective hot sedimentary aquifers (HSA) are naturally highly permeable, are overlain by rocks that act as thermal insulators, and are underlain by an effective source of heat energy (magma or high-heat-producing rocks such as granite plutons rich in uranium).

Definition of Geothermal Energy

Geothermal energy is the terrestrial generated heat stored in, or discharged from rocks and fluids (water, brines, gasses) saturated pore space, fractures, and cavities and is widely harnessed in two ways: for power (electricity) generation and for direct use, e.g., heating, cooling, aquaculture, horticulture, spas, and a variety of industrial processes, including drying. Thermal energy is used by taking heat from geothermal reservoirs replenished by natural recharge. Reservoirs that are naturally sufficiently hot and permeable are called hydrothermal reservoirs, whereas reservoirs that are sufficiently hot but require artificial improvement of a rock permeability are called engineered (enhanced) geothermal systems (EGS). Geothermal energy can be used to generate electricity or directly for processes that need thermal energy. Geothermal energy can be used to provide dispatchable, base-load electricity power plants.

Introduction

Geothermal energy systems have a modest environmental footprint, will not be impacted by climate change, and have potential to become the world's lowest cost source of sustainable thermal fuel for zero emission, base-load direct use, and power generation. Displacement of more emissive fossil energy supplies with geothermal energy can also be expected to play a key role in climate change mitigation strategies.

The use of energy extracted from temperatures of the Earth at shallow depth by means of ground source heat pumps (GSHP) is a common form of geothermal energy use. The direct uses of natural flows of

geothermally heated waters to surface have been practiced at least since the Middle Paleolithic [1], and industrial utilization began in Italy by exploiting boric acid from the geothermal zone of Larderello, where in 1904 the first kilowatts of electric energy (kWe) were generated and in 1913 the first 250-kWe commercial geothermal power plant was installed [2].

Where very high-temperature fluids ($>180^{\circ}\text{C}$) flow naturally to surface (e.g., where heat transfer by convection dominates), geothermal resources are the manifestation of two factors:

- A geologic heat source to replenish thermal energy outflow
- A hydrothermal reservoir that can be tapped to produce geothermal fluids for its direct use and/or for generating electricity

Elsewhere, a third geologic factor, the insulating capacity of rocks (acting as a thermal blanket) is an additional necessary natural ingredient in the process of accumulating usable, stored heat energy in geologic reservoirs that can be tapped to flow heat energy and replenished by convective and conductive *heat flow* from sources of geothermal energy.

Usable geothermal systems occur in a variety of geological settings. These are frequently categorized as follows:

1. High-temperature ($>180^{\circ}\text{C}$) systems at depths above (approximately) 3.5 km are generally associated with recent volcanic activity and mantle hot spot anomalies. Other high-temperature geothermal systems below (approximately) 3.5 km are associated with anomalously high-heat-producing crustal rocks, mostly granites.
2. Intermediate-temperature systems ($100\text{--}180^{\circ}\text{C}$).
3. Low-temperature ($<100^{\circ}\text{C}$) systems.

Both intermediate- and low-temperature systems are also found in continental settings, formed by above normal heat production through radioactive isotope decay; they include aquifers charged by water heated through circulation along deeply penetrating fault zones. However, there are several notable exceptions to these temperature-defined categories, and under appropriate conditions, high-, intermediate-, and low-temperature geothermal fields can be utilized for both power generation and the direct use of heat. Offshore

geothermal resources are also sometimes included in lists of ocean energy systems [3].

Geothermal systems can also be classified as: *convection-dominated systems*, which include liquid- and vapor-dominated hydrothermal systems; *conduction-dominated systems* which include hot rocks; and *hybrid systems* that are sourced from convection, conduction, and high-heat-producing source rocks. Geologic aquifers that overlie radiating sources of heat and gain heat via convection and/or conduction are sometimes called *hot sedimentary aquifer systems*.

The most widely recognized manifestations of geothermal energy are related to convective heat flow, including: hot springs and geysers (e.g., the movement of hot water to land surface); volcanoes (e.g., the movement of magma to land surface and sea floors); and certain forms of economically significant minerals deposits resulting from their recovery from the injection of geothermally heated fluids into lower temperature levels where minerals crystallize and are accumulated.

Geothermal wells produce naturally hot fluids contained in hydrothermal reservoirs from a continuous spectrum of natural high to low permeability and porosity (including natural fractures). The capacity of geothermal reservoirs to flow hot fluids can be enhanced with hydraulic fracture stimulation and chemical treatment (ex. acidization), creating artificial fluid pathways in *enhanced or engineered geothermal systems* (EGS) as well described in detail in Reference [4]. Once at surface, heated fluids can be used to generate electric energy in a thermal power plant, or used in other applications requiring heat, as heating and cooling of buildings, district heating systems, aquaculture, agriculture, balneology, industrial processes, and mineral drying. Space heating and cooling can also be achieved with GHP systems.

The number, depth, and diameter of geothermal energy production wells vary with local requirements for direct use and electricity power plants. Higher temperatures and higher flow rates result in more thermal energy production per well. Wells drilled to depths down to 3.5 km in volcanic areas frequently produce high-temperature ($>180^{\circ}\text{C}$) fluids to surface. Indeed, temperatures above $1,000^{\circ}\text{C}$ can occur at less than 10 km depth in areas of magma intrusion. Given the global average land area surface temperature of (about) 15°C and an approximate global geothermal

gradient for land areas outside volcanic settings of (about) $30^{\circ}\text{C}/\text{km}$, the same high temperature ($>180^{\circ}\text{C}$) can be reached (on average) at a depth of about 5.5 to 10 km below ground level.

Electricity Generation

The main types of geothermal power plants use direct steam (often called dry steam), flashed steam, and binary cycles.

Power plants that use dry and/or flashed steam to spin turbines are the most commonly deployed form of geothermal electricity generation. These plants use the heat energy contained in water and steam flowed from geothermal wells to spin turbines, converting thermal and kinetic energy to electrical energy.

Organic Rankine power plants employing secondary working fluids are increasingly being used for geothermal power generation. These so-called binary closed-loop power plants do not flow produced geothermal fluids directly into turbines. Thermal energy contained in water and/or steam produced from geothermal wells is transferred to a secondary working fluid using a heat exchanger (hence the term binary closed loop). Organic compounds with lower boiling points than water (such as isopentane that boils at atmospheric pressure at about 28°C) are often used as working fluids. The heat energy in the geothermal fluid boils the working fluid changing it from a liquid to a pressurized organic vapor within the closed loop, which can then be expanded in a turbine to spin a generator. The exhausted working fluid is cooled, condensed back into a liquid, pressurized, and then recycled into the heat exchanger to complete the cycle.

Direct Use

Direct use provides heating and cooling for buildings including district heating, fish ponds, greenhouses, bathing, wellness and swimming pools, water purification/desalination, and industrial and process heat for agricultural products and mineral extraction and drying.

For space heating, two basic types of systems are used: open or closed loop. Open loop (single pipe) systems utilize directly the geothermal water extracted from a well to circulate through radiators. Closed loop (double pipe) systems use heat exchangers to transfer heat from the geothermal water to a closed loop that

circulates heated freshwater through the radiators. This system is commonly used because of the chemical composition of the geothermal water. In both cases the spent geothermal water is disposed of into injection wells and a conventional backup boiler may be provided to meet peak demand.

Transmission pipelines for the direct use of geothermal energy consist mostly of steel insulated by rock wool (surface pipes) or polyurethane (subsurface). However, several small villages and farming communities have successfully used plastic pipes (polybutylene) with polyurethane insulation as transmission pipes in Iceland. It is for your consideration, as Iceland is mentioned below. The temperature drop is insignificant in large-diameter pipes with a high flow rate, as observed in Iceland where geothermal water is transported up to 63 km from the geothermal fields to towns.

It is debatable whether geothermal heat pumps (GHP), also called ground source heat pumps (GSHP), are purely an application of geothermal energy or also partially use stored solar energy. GHP technology is based on the relatively constant ground or groundwater temperature ranging from 4°C to 30°C to provide space heating, cooling, and domestic hot water for all types of buildings. Extracting energy during heating periods cools the ground locally. This effect can be minimized by dimensioning the number and depth of probes in order to avoid harmful impacts on the ground. These impacts are also reduced by storing heat underground during cooling periods in the summer months.

There are two main types of GHP systems: closed loop and open loop. In ground-coupled systems, a closed loop of plastic pipe is placed into the ground, either horizontally at 1–2 m depth or vertically in a borehole down to 50–250 m depth. A water-antifreeze solution is circulated through the pipe. Heat is collected from the ground in the winter and reinjected to the ground in the summer. An open-loop system uses groundwater or lake water directly as a heat source in a heat exchanger and then discharges it into another well or into the same water reservoir [5].

Heat pumps operate similarly to vapor-compression refrigeration units with heat rejected in the condenser for heating or extracted in the evaporator used for cooling. GHP efficiency is described by a coefficient of performance (COP) which scales the heating or cooling

output to the electrical energy input. GHPs typically exhibit between three and five COP [5, 6]. The seasonal performance factor (SPF) provides a metric of the overall annual efficiency of a GHP system. It is the ratio of useful heat to the consumed driving energy (both in kilowatt hour per year), and it is slightly lower than the COP.

Comparative Advantages of Geothermal Energy Use

Geothermal energy use has several comparative advantages in competitive energy markets.

- Geothermal plants have low-emission to emission-free operations and relatively modest land footprints. The average direct emissions yield of partially open cycle, hydrothermal flash, and direct steam electric power plants yield is about 120 g CO₂/kWh. This is the weighted average of 85% of the world's power plant capacity, according to References [7, 8]. Current binary cycle plants with total reinjection yield less than 1 g CO₂/kWh in direct emissions. Emissions from direct use applications are even lower [9]. Over its full life cycle (including the manufacture and transport of materials and equipment), CO₂ equivalent emissions range from 23 to 80 g/kWh for binary plants (based on References [10, 11]) and from 14 to 202 g/kWh for district heating systems and GHPs (based on Reference [12]). This means geothermal resources are environmentally advantageous and the net energy supplied more than offsets the environmental impacts of human, energy, and material inputs.
- Geothermal electric power plants have characteristically high-capacity factors; the average for power generation in 2009 is 74.5% (67,246 GWh_{electrical} used from installed capacity of 10.340 GW_{electrical} in December 2008 based on Reference [13]), and modern geothermal power plants exhibit capacity factors greater than 90%. This makes geothermal energy well suited for base-load (24/7), dispatchable energy use.
- The average estimated 27.5% capacity factor for direct use in 2009 (121.7 TWh_{thermal} used from installed capacity of 50.6 GW_{thermal} based on Reference [14]) can be improved with smart grids (as for domestic and industrial solar energy generation) by

employing combined heat and power systems, by using geothermal heat absorptive and vapor-compression cooling technology, and by expanding the distributed use of geothermal (ground source) heat pump for both heating and cooling applications, and

- Properly managed geothermal reservoir systems are sustainable for very long-term operation, comparable to or exceeding the foreseeable design life of associated surface plant and equipment.
- Displacement of more emissive fossil energy supplies with geothermal energy can also be expected to play a key role in climate change mitigation strategies.

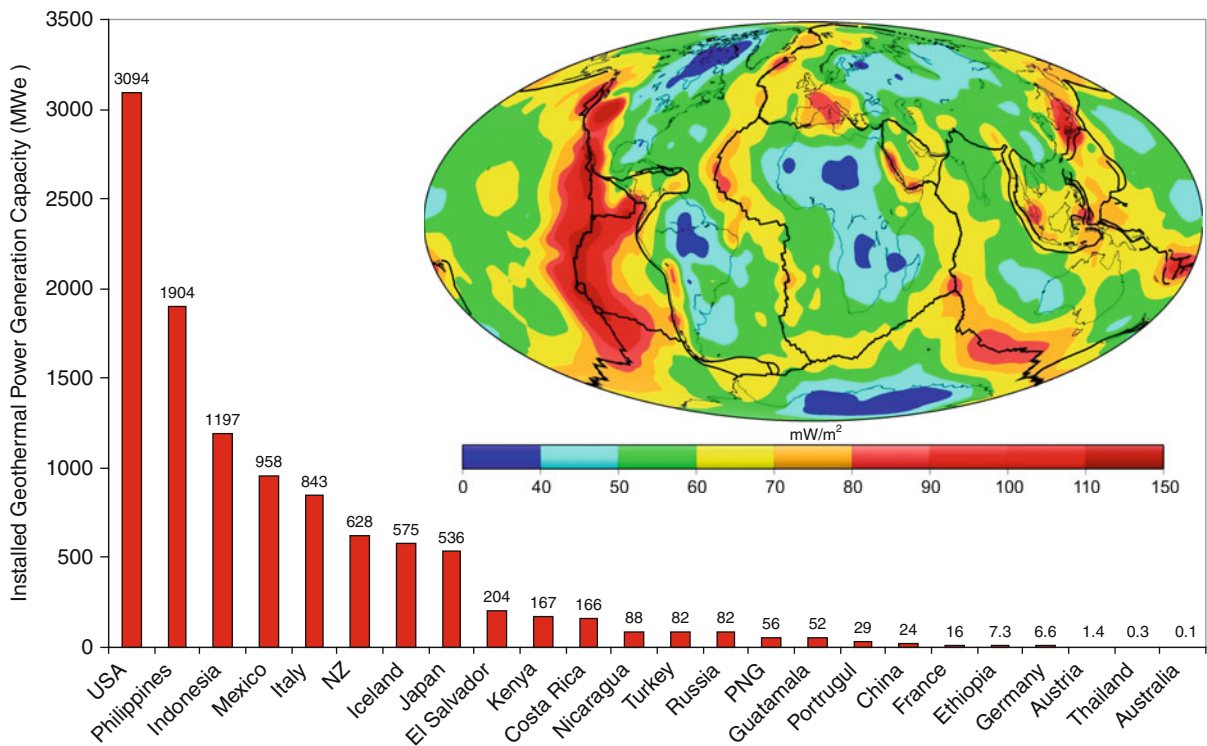
Geothermal Resources, Reserves, and Supplies

The theoretical global geothermal resource base corresponds to the thermal energy stored in the Earth's crust (heat in place). The technical (prospective) global geothermal resource is the fraction of the earth's stored heat that is accessible and extractable for use with foreseeable technologies, without regard to economics. Technical resources can be subdivided into three categories in order of increasing geological confidence: inferred, indicated, and measured [15] with measured geothermal resources evidenced with subsurface information to demonstrate its usability. Geothermal reserves are the portion of geothermal resources that can confidently be used for economic purposes. Geothermal reserves developed and connected to markets are energy supplies and global supplies.

Geothermal Supplies

At year-end 2010, geothermal energy supplies were used to generate base-load electricity in 24 countries with an installed capacity of nearly 11 GW of electricity and a global average capacity factor of nearly 75%, with newer installations above 90%, providing 10–30% of their electricity demand in six countries [13]. Figure 1 provides the geothermal electricity generation capacity by country and the mapped (estimated) distribution of global heat flow in milliwatts per square meter (mW/m²).

At year-end 2010, geothermal energy supplies are also used for direct use applications in 78 countries, accounting for 50 GW_{thermal} including district (space) heating and cooling and geothermal (ground source) heat pumps, which have achieved significant market



Geothermal Energy, Nature, Use, and Expectations. Figure 1

Geothermal electric installed capacity by country in 2009. This figure also depicts global average heat flow in milliwatts per square meter and tectonic plate boundaries (*black lines*) (Illustration adapted from a figure in Reference [16] with data from Reference [13]). This map of heat flow does not reconcile all geothermal information. The delineation of geothermal resources will be improved by integrating temperature gradient, heat flow, and reservoir data

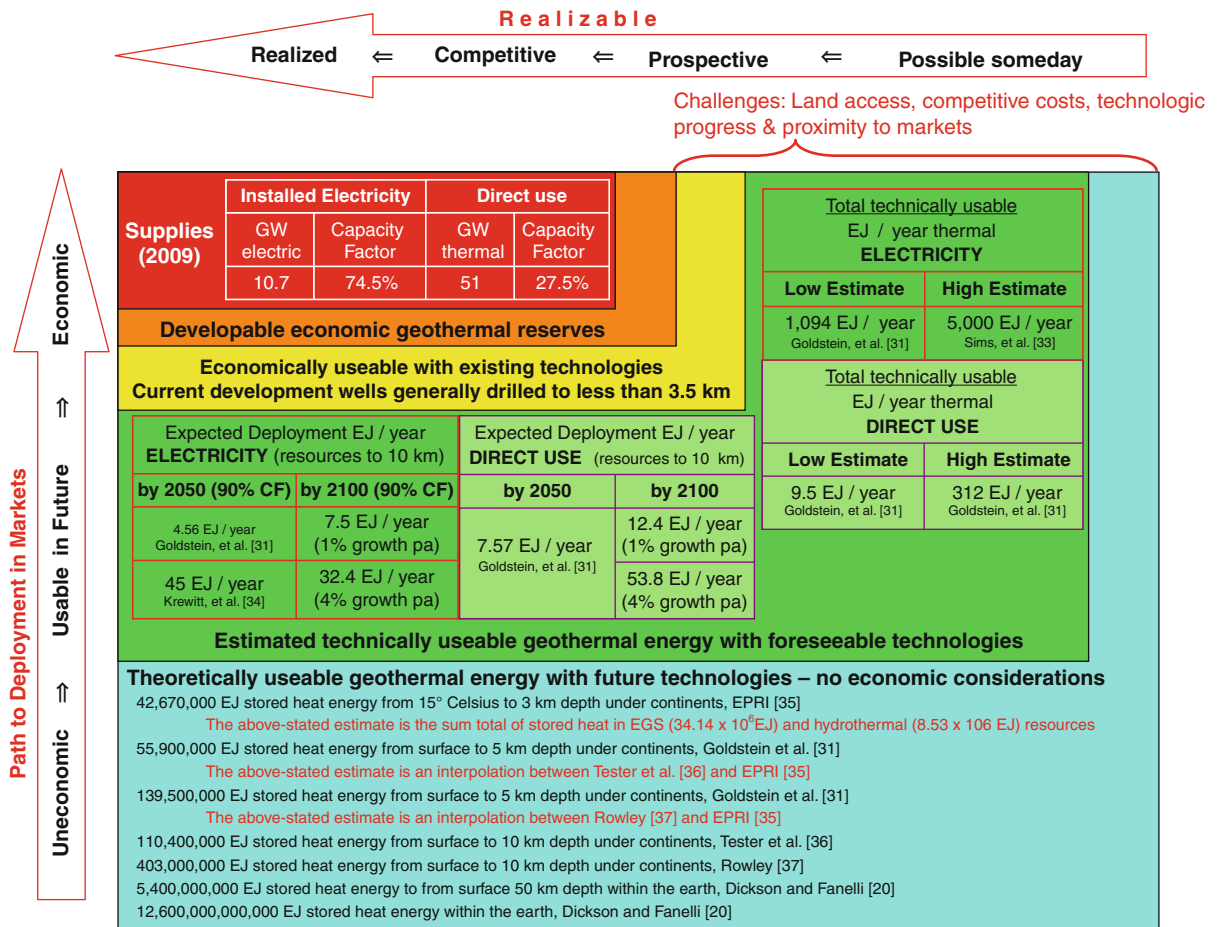
penetration worldwide [14]. Geothermal electric installed capacity by country in 2009. In the 40-year term 1970–2009, the average annual growth of geothermal electric installed capacity is 7% per annum; and in the 35-year term 1975–2009, the average annual growth for geothermal direct use is 11% per annum [13, 14, 17–19].

Geothermal Resources and Reserves

The total thermal energy contained in the Earth is on the order of 12.6×10^{12} EJ and that of the crust is on the order of 5.4×10^9 EJ to depths of up to 50 km [20]. The main sources of this energy are due to the heat flow from the earth's core and mantle and that generated by the continuous decay of radioactive isotopes in the crust itself. Heat is transferred from the interior toward the surface, mostly by conduction,

at an average of 0.065 W/m^2 on continents and 0.101 W/m^2 through the ocean floor. The result is a global terrestrial heat flow rate of around 1,400 EJ/year. Considering that continents cover $\sim 30\%$ of the earth's surface and their lower average heat flow, the terrestrial heat flow under continents has been estimated at 315 EJ/year [21].

Under continents, the stored thermal energy within 50, 10, 5, and 3 km depth (all depths reachable with the current drilling technology) has been estimated as presented as the theoretical usable geothermal energy in Fig. 2. For the Australian continent alone, Reference [24] estimated that recovery of just 1% of the stored geothermal energy above 150°C to 5 km in the Australian continental crust corresponds to 190,000 EJ. Based on these estimates, the theoretically available resource is enormous and clearly not a limiting factor for global geothermal deployment.



Geothermal Energy, Nature, Use, and Expectations. Figure 2

Potential geothermal energy resources split into categories, e.g., theoretical, technical, economic, developable, and existing supplies for power generation and direct use. All categories for power generation assume a 74.5% capacity factor and 8.1% average efficiency for converting thermal into electrical energy, though both factors will likely improve (increase) in future. All direct use estimates for the future assume an average 31% capacity factor, somewhat higher than the average (27.5%) in 2009 (Adapted from Fig. 1 in Reference [22] and the presentation by L. Rybach published in Reference [23])

Geothermal energy is a renewable resource. As thermal energy is extracted from the active reservoir, it creates locally cooler regions temporarily. Geothermal projects are typically operated at production rates that cause local declines in pressure and/or in temperature over the economic lifetime of the installed facilities. These cooler and lower pressure zones in the reservoir lead to gradients that result in continuous recharge by conduction from hotter rock and convection and advection of fluid from surrounding regions. Detailed modeling studies

[25, 26] have shown that resource exploitation can be economically feasible and still be renewable on a reasonable timescale when nonproductive recovery periods are considered.

Future Directions for Geothermal Energy Technologies

Challenges

Geothermal resources contain thermal energy that can be produced, stored, and exchanged (flowed) in rock,

gas (steam), and liquids (mostly water) in the subsurface of the earth.

With proper management practice, geothermal resources are sustainable and renewable over reasonable time periods. As stored thermal energy is extracted from local regions in an active reservoir, it is continuously restored by natural conduction and convection from surrounding hotter regions, and the extracted geothermal fluids are replenished by natural recharge and by reinjection of the exhausted fluids.

The obvious generalized impediments to massive, global geothermal energy use are:

- Currently insufficiently predictable reliability of geothermal reservoir performance (and in particular, the predictable reliability of engineered geothermal system reservoirs)
- Current costs of geothermal well deliverability (and, in particular, fluid production levels from stimulated engineered geothermal systems and the high costs of drilling deep wells)

Hence, the overarching common and well-justified objectives of global government initiatives are to stimulate technologic and learn-while-doing breakthroughs that will lead to a point where the cost of geothermal energy use is reliably cost competitive and comparatively advantageous within markets.

Priorities to Wider Use of Geothermal Energy

Improved, evermore reliable, cost-effective methods to enhance the productivity of geothermal systems will be essential to the competitiveness of geothermal resource in energy markets. In particular, the commercialization of fracture and/or chemical stimulation methods to reliably create engineered geothermal systems (EGS) independent of site conditions will be one key milestone on the road to great expectations for widespread economic use of geothermal energy. Table 1 results from a scan of the objectives of international geothermal energy fora and defines the top 20 priorities for advancing efficiency and competitiveness in geothermal energy use. This is an update of the priorities presented in Reference [27].

Geothermal Energy, Nature, Use, and Expectations.

Table 1 Top 20 research and development priorities for advancing efficiency and competitiveness in geothermal energy use

Openness to cooperation to engender complementary research and the sharing of knowledge	Informing industry, governments, and the public of technologic advances and the merits of using geothermal energy through presentations, publications, websites, submissions to enquiries, and the convening of conferences, workshops, and courses
Creating effective standards for reporting geothermal operations, resources, and reserves	For EGS, improved hard rock drill equipment
Predictive reservoir performance modeling	Improved multiple zone isolation for high-temperature and high-pressure geothermal reservoirs
Predictive stress field characterization	For deep geothermal reservoirs, reliable submersible pumps
For EGS, mitigate induced seismicity	Longevity of well cementing and casing
Condensers for high ambient surface temperatures	For EGS, optimum fracture stimulation methods
Use of CO ₂ as a circulating fluid	High temperature logging tools and sensors
Improve power plant design	High temperature flow survey tools
Technologies and methods to minimize water use	High temperature fluid flow tracers
Predict heat flow and reservoirs ahead of the drill bit	Mitigation of formation damage, scale, and corrosion

Expectations for Geothermal Energy Use

The extent or accessibility of geothermal resources will not be a limiting factor for deployment. The key determining factor in the growth in deployment will be the

Geothermal Energy, Nature, Use, and Expectations. Table 2 Global forecasts of: installed capacity for geothermal power generation (GWe), installed capacity to deliver thermal energy for direct use (GWt), geothermal power use (TWh_e/year), and geothermal direct uses (TWh_t/year)

Expected world use	2020		2030		2050		2100	
	Direct (GWt)	Electric (GWe)	Direct (GWt)	Electric (GWe)	Direct (GWt)	Electric (GWe)	Direct (GWt)	Electric (GWe)
Capacity	160.5	25.9	455.9	51.0	800	160.6	1,316–5,685	264–1,141
Expected global use	TWh_t/year	TWh_e/year	TWh_t/year	TWh_e/year	TWh_t/year	TWh_e/year	TWh_t/year	TWh_e/year
	421.9	181.8	1,998.8	380.0	2102.2	1266.4	3,457–14,940	2,083–9,000
	EJ/year	EJ/year	EJ/year	EJ/year	EJ/year	EJ/year	EJ/year	EJ/year
	1.52	0.65	4.41	1.37	7.57	4.56	12.4 to 53.8	7.5 to 32.4

Geothermal Energy, Nature, Use, and Expectations.

Table 3 Actual (from 1995 to 2010) and expected (from 2015 to 2100) growth in the use of geothermal energy

Year	Installed capacity actual or mean forecast (GWe)	Electricity production actual or mean forecast (GWh/year)	Capacity factor (%)
1995	6.8	38,035	64
2000	8.0	49,261	71
2005	8.9	56,786	73
2010	10.7	67,246	75
2015	18.5	121,600	77
2020	25.9	181,800	80
2030	51.0	380,000	85
2040	90.5	698,000	88
2050	160.6	1,266,400	90
2100	264–1,141	2,082,762–8,999,904	90+

competitiveness of geothermal energy use within local, regional, national, and trade zone markets. The authors have drawn conclusions in regard to future growth in the use of geothermal energy through 2010. The following table (Table 2) provides those global long-term forecasts of installed capacity for geothermal power and direct uses (heat) and of electric and direct

uses (heat) generation. Earlier estimates for deployment beyond 2010 that were considered in developing forecasts include References [13, 28–30].

The above-listed forecasts assume improvements in capacity factors power generation from the current average 74.5–90% by 2050, a level already attained in efficient, existing geothermal power generation plants. A more detailed account of actual and expected growth in the use of geothermal energy follows (Table 3). The statistics for installed capacity to generate electricity from geothermal energy, electricity production from those geothermal plants, and capacity factors for geothermal power plants are from the Reference [29] for the term 1995–2005; Reference [13] for 2010 and the Reference [31] for the term 2015–2100. The expressed forecasts for growth from 2050 are based on 1% and 4% average annual growth for the 50 years to 2100.

Next Steps in Global Resource Assessments

A further global geothermal resource assessment is planned under an existing IEA Geothermal Implementing Agreement research annex. This will include a probabilistic range of estimates, e.g., assuming that a log-normal distribution adequately describes the range of recovery of stored heat from a minimum of 0.5% at a 99% probability to a maximum of 40% of stored heat at a 1% probability. This implies: a low-side recovery of 1.34% of stored heat (90% probability), a mid-range recovery of 4.47% of stored heat

(50% probability), a Swanson's mean recovery of 6.68% of stored heat, and a high-side recovery of 14.95% of stored heat (10% probability). (Swanson's mean is the weighted approximation for a log-normal distribution equal to the summation of 30% of the 90% probability value, 30% of the 10% probability value, and 40% of the 50% probability value, e.g., $(P90 \times 0.3) + (P10 \times 0.3) + (P50 \times 0.4)$ equals the Swanson's mean value.)

Key Points

- With its natural thermal storage capacity, geothermal is especially suitable for supplying both base-load electric power generation and for fully dispatchable heating and cooling applications in buildings, and thus is uniquely positioned to play a key role in climate change mitigation strategies [32].
- Direct use of geothermal energy for heating and cooling, including geothermal heat pumps (GHPs), is expected to increase to 7.86 EJ/year (~ 815 GWt) by 2050 and between 12.9 EJ/year (with 1% growth per year) and 55.9 EJ/year (with 4% growth per year) by 2100. Marketing and multiple internationally competitive supply chains will underpin this growth. This expectation is supported information published by Reference [6].
- Power generation with binary plants and total reinjection will become commonplace in countries without high-temperature resources.
- Geothermal energy utilization from conventional hydrothermal resources continues to accelerate, and the advent of EGS is expected to rapidly increase growth after 10–15 years putting geothermal on the path to provide an expected generation global supply of 4.56 EJ/year (~ 160 GWe) by 2050 and between 7.5 EJ/year (with 1% growth per year) and 32.4 EJ/year (with 4% growth per year) by 2100.
- Geothermal energy is expected to meet between 2.5% and 4.1% of the total global demand for electricity by 2050 and potentially more than 10% by 2100. It is also expected to provide about 5% of the global demand for heating and cooling by 2050 and, potentially, more than 10% by 2100. Geothermal energy will be a dominant source of base-load renewable energy in many countries in the next century.
- In addition to the widespread deployment of EGS, the practicality of using supercritical temperatures and offshore resources is expected to be tested with experimental deployment of one or both a possibility by 2100.

Acknowledgments

The authors thank their international colleagues who have contributed so much of their professional lives and time to provide improved understanding of geothermal systems. We are especially grateful to Ken Williamson, David Newell, Trevor Demayo, Arthur Lee, Subir Sanyal, Roland Horne, David Blackwell, Greame Beardsmore, and Doone Wyborn.

Bibliography

Primary Literature

1. Cataldi R (1999) The year zero of geothermics. In: Cataldi R, Hodgson S, Lund JW (eds) *Stories from a heated earth*. Geothermal Resources Council, Sacramento, pp 7–17. ISBN 0934412197
2. Burgassi PD (1999) Historical outline of geothermal technology in the Larderello region to the middle of the 20th century. In: Cataldi R, Hodgson S, Lund JW (eds) *Stories from a heated earth*. Geothermal Resources Council, Sacramento, pp 195–219. ISBN 0934412197
3. Hiriart G, Prol-Ledesma RM, Alcocer S, Espíndola G (2010) Submarine geothermics: hydrothermal vents and electricity generation. In: *Proceedings world geothermal congress 2010, Bali, Indonesia, 25–29 April 2010*
4. Tester JW, Anderson BJ, Batchelor AS, Blackwell DD, DiPippo R, Drake EM (eds) (2006) *The future of geothermal energy impact of enhanced geothermal systems on the United States in the 21st century*. Prepared by the Massachusetts Institute of Technology, under Idaho National Laboratory subcontract no. 63 00019 for the U.S. Department of Energy, Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Geothermal Technologies, p 358. ISBN-10: 0486477711, ISBN-13: 978-0486477718
5. Lund JW, Sanner B, Rybach L, Curtis R, Hellström G (2003) Ground-source heat pumps – A world overview. *Renew Energy World* 6(14):218–227, ISSN 1462-6381
6. Rybach L (2005) The advance of geothermal heat pumps world-wide. *IEA Heat Pump Centre Newsletter* 23, pp 13–18. ISSN: 0724-7028
7. Bertani R, Thain I (2002) Geothermal power generating plant CO₂ emission survey. *IGA News* 49, pp 1–3. ISSN: 0160-7782

8. Bloomfield KK, Moore JN, Neilson RN (2003) Geothermal energy reduces greenhouse gases. *Geoth Res Counc Bull* 32(2):77–79, ISSN 01607782
9. Fridleifsson IB, Bertani R, Huenges E, Lund JW, Ragnarsson A, Rybach L (2008) The possible role and contribution of geothermal energy to the mitigation of climate change. In: IPCC scoping meeting on renewable energy sources, Luebeck, Germany, 21–25 Jan 2008, p 36. Available at: <http://www.ipcc.ch/pdf/supporting-material/proc-renewables-luebeck.pdf>
10. Frick S, Schröder G, Kaltschmitt M (2010) Life cycle analysis of geothermal binary power plants using enhanced low temperature reservoirs. *Energy* 35(5):2281–2294, ISSN: 0360–5442
11. Nill M (2004) Die zukünftige Entwicklung von Stromerzeugungstechniken, Eine ökologische Analyse vor dem Hintergrund technischer und ökonomischer Zusammenhänge, Fortschritt-Berichte VDI Nr. 518. VDI, Düsseldorf, p346, ISSN: 0178–9414
12. Kaltschmitt M (2000) Environmental effects of heat provision from geothermal energy in comparison to other resources of energy. In: *Proceedings world geothermal congress 2000*, Kyushu-Tohoku, Japan, 28 May–10 Jun 2000. ISBN: 0473068117
13. Bertani R (2010) World update on geothermal electric power generation 2005–2009. In: *Proceedings world geothermal congress 2010*, Bali, Indonesia, 25–30 Apr 2010
14. Lund JW, Freeston DH, Boyd TL (2010) Direct utilization of geothermal energy 2010 worldwide review. In: *Proceedings world geothermal congress 2010*, Bali, Indonesia, 25–30 Apr 2010
15. AGEAG-AGEA (2009) Australian code for reporting of exploration results, geothermal resources and geothermal reserves, prepared by the Australian Geothermal Code Committee – A committee of the Australian Geothermal Energy Group (AGEG) and the Australian Geothermal Energy Association (AGEA). Download from: http://www.pir.sa.gov.au/geothermal/ageg/geothermal_reporting_code
16. Hamza VM, Cardoso R, Ponte Neto C (2008) Spherical harmonic analysis of earth's conductive heat flow. *International Journal of Earth Sciences* 97(2):205–226, (22), Springer
17. Lund JW, Freeston DH, Boyd TL (2005) Direct application of geothermal energy: 2005 worldwide review. *Geothermics* 34:691–727, ISSN 0375–6505
18. Garwell K, Greenberg G (2007) 2007 Interim report. Update on world geothermal development. Publication of the Geothermal Energy Association. Available at the GEA website: <http://www.geo-energy.org/reports/GEA%20World%20Update%202007.pdf>
19. Fridleifsson IB, Ragnarsson A (2007) Geothermal energy. In: 2007 survey of energy resources, World Energy Council 2007, pp 427–437. ISBN: 0946121 26 5. Available at: http://www.worldenergy.org/documents/ser2007_final_online_version_1.pdf
20. Dickson MH, Fanelli M (2003) Geothermal energy: utilization and technology. UNESCO, Renewable Energy Series, New York, p 206. ISBN 978-92-3-103915-7
21. Stefansson V (2005) World geothermal assessment. In: *Proceedings world geothermal congress 2005*, Antalya, Turkey, 24–29 April 2005. ISBN: 9759833204
22. Rybach L (2010) “The future of geothermal energy” and its challenges. In: *Proceedings world geothermal congress 2010*, Bali, Indonesia, 25–29 Apr 2010
23. Mongillo MA (2010) Proceedings of the Joint GIA-IGA workshop – Geothermal energy global development potential and contribution to mitigation of climate change, Madrid, Spain, 5–6 May 2009. Download from: http://www.iea-gia.org/documents/ProcGIA_IGAWorkshopMadridFinalprepress22Mar10_000.pdf
24. Budd AR, Holgate FL, Gerner E, Ayling BF, Barnicoat A (2008) Pre-competitive geoscience for geothermal exploration and development in Australia: geoscience Australia's onshore energy security program and the geothermal energy project. In: Gurgenci H, Budd AR (eds) *Proceedings of the Sir Mark Oliphant international frontiers of science and technology Australian geothermal energy conference*, Geoscience Australia, Record 2008/18. ISBN: 9781921498190
25. Pritchett R (1998) Modeling post-abandonment electrical capacity recovery for a two-phase geothermal reservoir. *Trans Geoth Resour Counc* 22:521–528, ISSN: 0193–5933
26. O'Sullivan M, Mannington W (2005) Renewability of the Wairakei-Tauhara geothermal resource. In: *Proceedings world geothermal congress 2005*, Antalya, Turkey, 24–29 April 2005. ISBN: 9759833204
27. Goldstein BA, Hill AJ, Long A, Budd AR, Holgate F, Malavazos M (2009) Hot rock geothermal energy plays in Australia. *Proceedings of the 34th workshop on geothermal reservoir engineering*, Stanford University, Stanford, California, 9–11 Feb 2009, SGP-TR-187
28. IPCC (2007) Climate change 2007, fourth assessment report of the intergovernmental panel on climate change, working group III: mitigation of climate change, chapter 4 – Geothermal, section 4.3.3.4
29. International Energy Agency (IEA) (2008) World energy outlook 2008. Download from: <http://www.iea.org/textbase/nppdf/free/2008/weo2008.pdf>
30. European Renewable Energy Council (EREC) and Greenpeace International (GPI) (2008) Energy [R]evolution – A sustainable global energy outlook (EREC-GPI-08). Download from: <http://www.greenpeace.org/raw/content/international/press/reports/energyrevolutionreport.pdf>
31. Goldstein BA, Hiriart G, Tester JW, Bertani R, Bromley CJ, Gutiérrez-Negrín LC, Huenges E, Ragnarsson A, Mongillo MA, Muraoka H, Zui VI (2011) Great expectations for geothermal energy to 2100. In: 36th Stanford workshop of geothermal reservoir engineering
32. Bromley CJ, Mongillo MA, Goldstein B, Hiriart G, Bertani R, Huenges E, Muraoka H, Ragnarsson A, Tester J, Zui V (2010) IPCC renewable energy report: the potential contribution of geothermal energy to climate change mitigation. In: *Proceedings world geothermal congress 2010*, Bali, Indonesia, 25–30 Apr 2010
33. Sims REH, Schock RN, Adegbulugbe A, Fenhann J, Konstantinavičiute I, Moomaw W, Nimir HB, Schlamadinger B, Torres-Martínez J, Turner C, Uchiyama Y, Vuori SJV, Wamukonya N,

- Zhang X (2007) Energy supply. In: B Metz, OR Davidson, PR Bosch, R Dave, LA Meyer (eds) Climate change 2007: mitigation. Contribution of working group III to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge/New York. ISBN-13: 9780521705981
34. Krewitt W, Nienhaus K, Klebmann C, Capone C, Stricker E, Grauss W, Hoggwijk M, Supersberger N, Von Winterfeld U, Samadi S (2009) Role and potential of renewable energy and energy efficiency for global energy supply. Climate change, 18 Dec 2009, p 344. ISSN: 1862–4359
 35. Electric Power Research Institute (EPRI) (1978) Geothermal energy prospects for the next 50 years. ER-611-SR, Special report for the world energy conference 1978
 36. Tester JW, Drake EM, Golay MW, Driscoll MJ, Peters WA (2005) Sustainable energy – Choosing among options. MIT Press, Cambridge, MA, p 850
 37. Rowley JC (1982) Worldwide geothermal resources. In: Edwards LM et al (eds) Handbook of geothermal energy. Gulf Publishing, Houston, pp 44–176, Chapter 2. ISBN 0-87201-322-7

Websites, Books and Reviews

- Beardmore GR, Cull JP (2001) Crustal heat flow: a guide to measurement and modelling. Cambridge University Press, Cambridge/New York
- DiPippo R (2008) Geothermal power plants. Principles, applications, case studies and environmental impact, 2nd edn. Elsevier, Burlington
- Huenges E (2010) Geothermal energy systems: exploration, development, and utilization. Wiley-VCH, Weinheim
- Enhanced Geothermal Innovative Network for Europe (ENGINE), <http://engine.brgm.fr/>
- European Energy Research Alliance Joint Programme on Geothermal Energy (EERA JPGE), <http://www.eera-set.eu/index.php?index=36>
- European Geothermal Energy Council (EGEC), <http://www.egec.org/>
- Geo-Heat Center (GHC), <http://geoheat.oit.edu/>
- Geothermal Education Office (GEO), <http://geothermal.marin.org/>
- Geothermal Engineering Integrating Mitigation of Induced Seismicity in Reservoirs (GEISER), <http://www.geiser-fp7.eu/default.aspx>
- Geothermal Resource Association (GEA), <http://www.geo-energy.org/>
- Geothermal Resource Council (GRC) and its annual conference in particular, <http://www.geothermal.org/>
- International Energy Agency's Geothermal Implementing Agreement (IEA GIA), <http://www.iea-gia.org/>
- International Geothermal Association (IGA) and its World Geothermal Congress (WGC) ^(a), <http://www.geothermal-energy.org/>
- International Panel for Climate Change (IPCC). Working group III – Special report on renewable energy (and in particular Chapter 4 – Geothermal), <http://www.ipcc-wg3.de/publications/special-reports/special-report-renewable-energy-sources>

- International Partnership for Geothermal Technologies (IPGT) ^(a), <http://internationalgeothermal.org/>
- Stanford University Geothermal Workshops, <http://pangea.stanford.edu/ERE/research/geoth/conference/workshop.html>
- US Department of Energy, <http://www.energy.gov/energysources/geothermal.htm>

Geothermal Field and Reservoir Monitoring

TREVOR M. HUNT

GNS Science, Wairakei Geothermal Research Centre, Taupo, New Zealand

Article Outline

Glossary
 Definition of the Subject and Its importance
 Introduction
 Purposes and Principles of Monitoring
 Down-Hole Monitoring
 Surface Monitoring
 Future Directions
 Bibliography

Glossary

- Anchor grouting** Concrete pumped into the rocks around the upper part of the well to anchor the well and well cellar to the near-surface rock formations.
- Aquiclude** A geological formation (or formations) which will not transmit water; a barrier to vertical movement of geothermal fluid.
- Aquifer** A geological formation (or formations) which contains water or geothermal fluid and will allow fluid movement.
- Baseline** Data set acquired before exploitation begins, against which any future measurements are compared.
- Benchmark** Permanent survey mark, often consisting of a stainless steel pin set in a concrete block or in the concrete base of a pipeline support.
- Bleed** A well that is throttled back to a minimum flow is said to be “on bleed.” It is often risky to completely shut down a geothermal production well because it may be difficult to restart. Bleeding also keeps the wellbore heated which minimizes corrosion.

- Deep liquid level** Boundary between the two-phase and deep liquid zones.
- Deep liquid zone** Region of single phase liquid conditions below a two-phase (liquid and vapor) zone.
- Developer** Company or organization which locates or uses geothermal energy for domestic or industrial purposes.
- Dryout** The process whereby liquid saturation in the pores decreases and the vapor saturation increases, as a result of a decrease in pressure.
- Epicenter** The point on the Earth's surface directly above the hypocenter or focus of an earthquake.
- Geothermal system** A body of hot water and rock within the Earth.
- Go-devil** A tool for determining wellbore clearances or for scraping out obstructions from a well or pipeline.
- Groundwater** Water, generally cold and of meteoric origin, which resides in near-surface aquifers and is often used for domestic and industrial purposes.
- High-temperature system** A geothermal system, or part thereof, containing fluid having a temperature greater than 150°C; c.f. *low-temperature system* in which the temperature is less than 150°C. Note, however, that this temperature value is arbitrary and that different authorities adopt different values, or divide the range into low, intermediate, and high temperature.
- Hypocenter** The focus or focal point of an earthquake (x, y, z) c.f. epicenter (x, y).
- Injection (syn. reinjection)** The process of returning waste water from a geothermal power station or industrial process back into the ground. This generally occurs around the edges of the field and may not be into the production aquifer from which fluid is drawn off to the power station.
- Injection aquifer** The formation into which injected fluid is put. Generally this has high porosity and permeability.
- Liquid-dominated system** A geothermal system, or part thereof, in which the pressure is hydrostatically controlled; c.f. *steam (vapor)-dominated system*, where the pressure is steam-static.
- Make-up well** Well drilled to replace production lost from an existing production well, due to decreases in fluid temperature or pressure.
- Perched aquifer** An aquifer of limited lateral extent which is separated from an underlying body of groundwater by unsaturated rock.
- Permeability** A measure of the capacity of a geological rock formation to transmit a fluid.
- Production zone** That region (depth) of the geothermal reservoir from which most of the production of fluid occurs.
- Reservoir** The region of a geothermal system from which geothermal fluid is withdrawn, or is capable of being withdrawn.
- Residual (liquid) saturation** The amount of liquid that remains in the pores (as % of pore volume) which decreases in pressure will not vaporize. The liquid saturation level below which vaporization of liquid will not occur.
- Steam zone** A region of the reservoir in which steam (vapor) is the pressure-controlling phase.
- Trigger point** A measured value at which it is considered action needs to be taken to prevent or avoid some detrimental occurrence happening, or exceeding some predetermined limit.
- Two (2)-phase zone** A region where the liquid and vapor (steam) phases of water coexist in pores or fractures.
- Vadose zone** The region of unsaturated rock and soil between the ground surface and the shallow groundwater level.
- Waste water** Geothermal water from which energy has been extracted and is no longer required. This may be separated water, or steam which has passed through turbines or a binary plant and been condensed.

Definition of the Subject and Its Importance

Geothermal systems are dynamic entities in which the liquid and vapor phases of water are the main mobile constituents. In their natural state these are generally in a quasi steady-state condition, when considered over a long period of time (>1,000 years). However, when fluid is withdrawn for the purpose of extracting energy then changes may occur within the system. These changes can result in a variety of environmental effects some of which are undesirable and so to manage the extraction of energy in a sustainable and environmentally responsible way it is necessary to monitor the

changes. By monitoring the changes with time it is possible to understand and model the effects these changes may have on the environment and take steps to minimize any undesirable effects in a timely manner. Changes may also have engineering implications for a geothermal development, especially for a power station. One example is a decrease in the pressure of steam supplied to the station that may necessitate replacement of the original turbines by those designed to operate at lower pressures. At the start of production at Wairakei (New Zealand) in 1958, the high-pressure (HP) turbine inlet pressure was 1.25 MPa, but by the late 1970s the pressure had fallen to about 0.7 MPa, and the HP turbines were taken offline and the wells derated to intermediate pressure [1]. Another example is a change in enthalpy due to a change in the steam-water ratio that may affect the efficiency of a modular binary plant designed for a specific steam-water mixture.

Introduction

In a typical high-temperature geothermal system used for electrical power generation, a large mass of hot water is withdrawn from an area and the cooler, waste water is injected in a different location, and this can give rise to significant changes within the system and at the surface. However, in low-temperature systems or where only heat (no mass) is extracted, the changes may be small and negligible.

Purposes and Principles of Monitoring

Purposes of Monitoring

Where significant changes occur, or it is anticipated they might occur, a developed geothermal system will be monitored to:

1. Obtain data on which rational and informed resource management decisions can be made by developers and regulatory authorities.
2. Verify that management decisions are having the desired outcomes.
3. Enable the public to have confidence in the environmental management process.
4. Assist in building up knowledge of geothermal systems and how to develop them in a sustainable and environmentally responsible way.

Basic Principles

Ideally, monitoring begins before development starts so that a good baseline is obtained. It is not possible to go back in time, so many different eventualities need to be considered and a fully integrated monitoring program needs to be developed and begun before large-scale productions starts.

Monitoring should be conducted at a frequency sufficient to enable natural variations to be distinguished from exploitation-induced changes.

The data collected needs to be interpreted and regularly compared with predetermined “trigger points.” No change may be as important as some change, and is not a valid reason for stopping monitoring, although the frequency of measurement may be reduced after a long period of no change.

Data need to be reliable. Equipment should be calibrated regularly and operated by a competent person. Since monitoring may continue over a long period of time, it is important that the same techniques are used such that a valid comparison can be made between early and recent data.

Monitoring Program Planning

A geothermal monitoring program is likely to extend for several decades, therefore all observations and measurements need to be carefully documented in a suitable archive. During this time there will probably be staff changes and therefore there needs to be a written set of instructions about how and when measurements will be made, so that measurements at different times are compatible with each other. Monitoring sites need to be clearly marked and monitoring facilities (e.g., groundwater monitor wells) need to be maintained. Experience has shown that large-scale geothermal developments often start as a small development and increase in size incrementally. Furthermore, as production wells decline new wells are drilled to maintain steam quantities deliverable to the power station. These engineering activities may result in monitoring sites being altered or destroyed, so it is important that the baseline data set has sufficient redundancy to allow for such loss without seriously compromising its integrity.

Interpretation of Monitoring Data

Generally the process of collecting monitoring data is relatively easy, however, correct interpretation of the results may be difficult. Often the first problem in interpretation is separating natural variations from those induced by utilization of the field. Further complexities may be introduced by other human activities, for example, pumping of water supply wells for irrigation or drinking water for animals, and diversion or damming of rivers may cause groundwater level changes. The effects of some anthropogenic changes may be difficult to measure or even estimate. Another significant problem is that what is measured inside a drillhole may not represent what is occurring in the rock outside the drillhole, because as fluid passes from the rock into the hole it may change in character.

Down-Hole Monitoring

A variety of down-hole monitoring techniques have been developed, many originating from the oil industry, to determine reservoir changes. Ideally, down-hole monitoring is undertaken in nonproducing wells, or production wells that are shut down or on “bleed.” However, making the measurements is the simplest part of the process because often the casing configuration of the well can strongly influence the data obtained and needs to be taken into account. Furthermore, there is a problem in that conditions within a wellbore may be different from those in the rock outside the wellbore.

Pressure

Changes in fluid pressure with depth and time are key indicators of reservoir changes, especially in high-temperature liquid-dominated systems. In their natural, predevelopment state the reservoirs in such systems contain boiling liquid water, with pockets of 2-phase conditions in the upper part. Pressures are near boiling point for depth. Extraction of fluid, and the concomitant decrease in pressure, generally causes these pockets of 2-phase conditions to coalesce into a continuous 2-phase zone and then expand (both horizontally and vertically). As the pressures fall the liquid saturation in the pores and fissures decreases, and eventually steam becomes the pressure-controlling phase in the upper part of the zone, giving rise to a steam zone in which

there is negligible change in pressure with depth (Fig. 1). Continued pressure decreases (e.g., Fig. 2) may lead to cool inflows which resaturate the pores in the 2-phase zone, resulting in a rise in the deep liquid level and an increase in pressure in the deep liquid zone (although pressures may continue to decrease in the overlying steam and 2-phase zones).

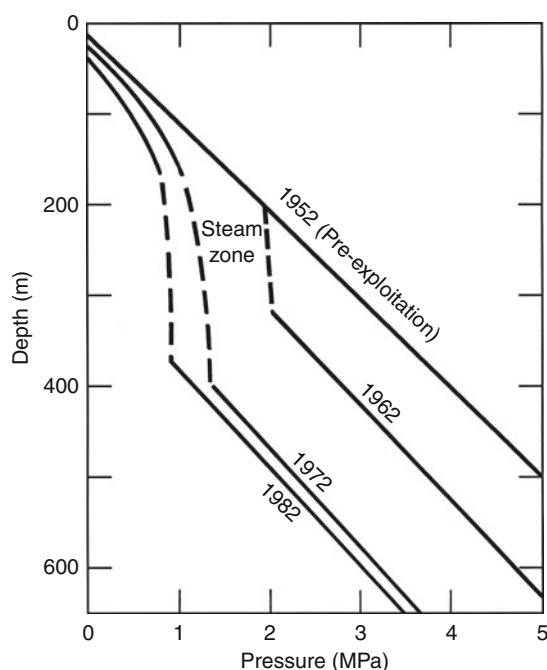
A continuous vertical profile of pressure variation with depth is made using high-pressure high-temperature (HPHT) wire line equipment, now generally using quartz pressure transducers.

Temperature

Changes in fluid temperature within the geothermal reservoir are of vital importance in managing and guiding future development of a geothermal field. After drilling and output testing, a geothermal well is generally left for several weeks for the temperatures of the wellbore fluid to stabilize and achieve thermal equilibrium with the surrounding rock formations. A profile of temperature variation with depth is then made using HPHT wire line equipment, and repeated at intervals of time. For production wells this temperature logging can only be done when it is possible to shut the well down for enough time for thermal equilibrium to occur.

Initial temperature logs of deep geothermal wells rarely show a consistent increase of temperature with depth: regions of cooler values (reversals) reveal cool inflows and regions of hotter values reveal feed zones (Fig. 3). When interpreting temperature logs it is important to take into account the casing pattern; in uncased regions or zones of slotted casing, there may be flows within the wellbore between different formations (see below) that result in the measured temperatures being different from the rock outside the wellbore.

Some production wells may experience significant decreases in feed temperatures with time as a result of cold downflows, lateral inflows of cooler water, or changes in the relative amounts of contribution from feed zones of different temperature (Figs. 4 and 5). Lateral inflows of cooler water may be associated with the return of cooler injected water along high permeability paths. Temperature and chemistry monitoring can detect such returns and so guide the location of drilling of make-up wells.



Geothermal Field and Reservoir Monitoring. Figure 1 Sketch showing the variation of pressure with depth at different times during early development of the Wairakei geothermal field, New Zealand (Taken from [2]). Note increase in thickness of the steam zone (1962, 1972) followed by slight decrease (1982)

Flow

Changes in fluid mass and volumetric flow rate, and the proportion of vapor (steam) to liquid (water), for individual production wells are also important for managing and guiding the development of a liquid-dominated geothermal field. Measuring the mass and volumetric flow in a well or pipeline is not easy, especially two-phase flow, and accuracies can vary [4]. One modern method, a vortex mass flowmeter, is based on the phenomenon of vortex shedding. A non-streamlined body, called a “shedder bar,” inside the pipeline causes an alternating series of vortices to be shed from each side of the body. The distance between successive vortices on each side is related to the fluid velocity, and is measured by a sensor behind the shedder bar. From the velocity measurement, and simultaneously measured temperature and pressure values, the volumetric and mass flow rates can be calculated. Flow

rates within a well can be measured using a spinner type or hot-wire flowmeter. Generally these measurements are made at low flow rates or when the well is shut down and the instrument is passed through a gland at the wellhead. Another type of flowmeter used is the Coriolis or Inertial flowmeter which directly measures the mass flow rate [5]. It has one or more bent, straight, or U-shaped vibrating tubes in the fluid stream, and as the fluid passes through the tubes, they twist. The amount of tube twisting is directly proportional to mass flow. This meter can also be used for heat measurement of low-pressure, superheated steam. The chief advantage of a Coriolis flowmeter is in providing highly accurate measurements of mass flow rate without flow conditioning or accessory devices such as pressure or temperature measurement.

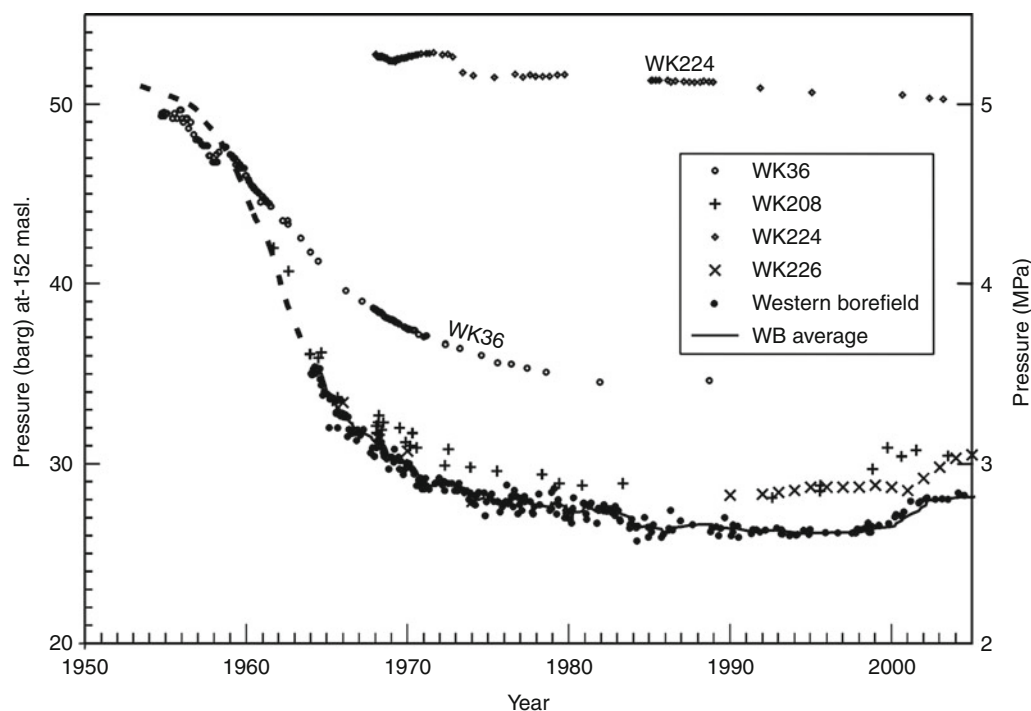
Fluid flow within a wellbore may be complex. In a flowing production well there may be several feed zones at different depths, each contributing to the total flow measured at the wellhead. In a non-flowing geothermal well, the high aspect ratio of the cased part of the wellbore precludes thermally generated fluid movement. However, in an open (uncased) or slotted region of the well there may be fluid movement between formations having different temperatures and physical properties; fluid may exit from one aquifer, travel up or down the wellbore, and enter another (thief zone). Repeated wire line flowmeter measurements may be made to detect changes in flow from the different feed zones over time, or before and after maintenance on a well, and for finding holes in casing that have developed due to corrosion.

Fluid flow within the rock matrix can be measured using chemical tracers introduced into a well and their arrival time and concentration measured in other wells (see below).

Generally, for a high-temperature geothermal field the mass flow rates will decline, and (for liquid-dominated wells) the enthalpy will increase over time, unless injection returns cause an increase in mass flow and decline in enthalpy.

Casing Integrity

Over a period of time, steel well casing and piping, and concrete anchor grouting can become damaged as a result of corrosion (stress or fatigue), scaling, or



Geothermal Field and Reservoir Monitoring. Figure 2

Changes in deep liquid pressure with time in Wairakei geothermal field (Taken from [3]). Pressure: 1 bar = 0.1 MPa. Testing of exploration wells began in the early 1950s and production started in 1958. Well WK224 (top) lies outside the field

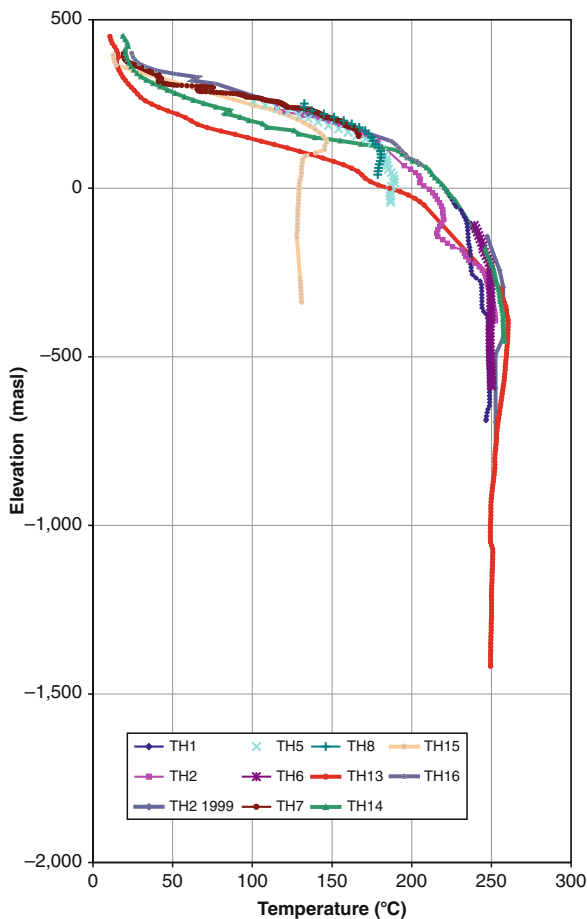
ground deformation. Corrosion may occur on the inner surface of the casing as a result of acidic fluids flowing up from deep feed zones, or it may occur on the outer surface of the casing and attack grouting as a result of acidic fluids being present in a formation through which the well passes [6, 7]. Corrosion is present in most geothermal wells, even low-temperature and low-enthalpy wells, but is generally more of a problem in deep, high-temperature, and high-enthalpy wells tapping CO_2 -rich and acidic fluids. Mechanical deformation (breaks, buckling) may occur in wells subject to significant ground deformation [8].

Casing damage is detected by running a mechanical caliper tool with flexible fingers up and down the hole, or by running a “Go-devil” tool down the hole. Detailed determination of the damage may then be investigated using a video camera or a sonic borehole televiewer. Damage to near-surface (<10 m depth) casing can be repaired by excavating a pit around the well and replacing the damaged casing. To repair deeper damage it may be necessary to run new

liner of smaller diameter inside the damaged production casing, apply casing patches or install expandable casing.

Fluid Chemistry

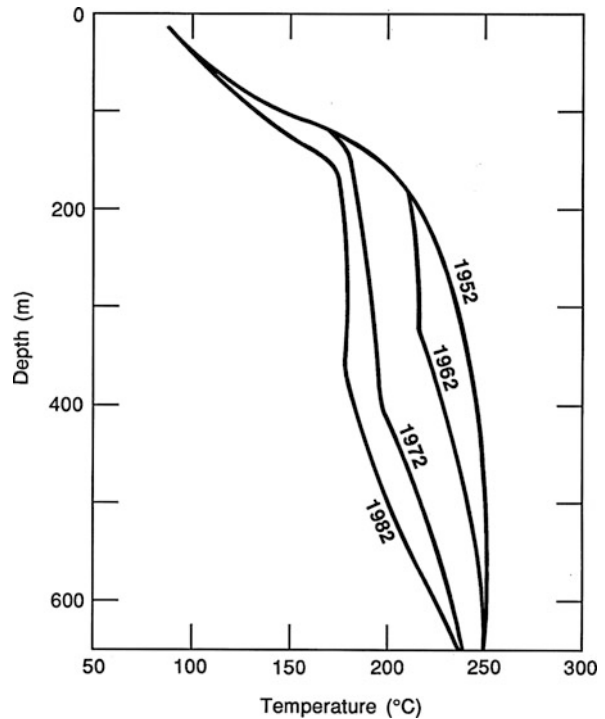
The chemical composition of geothermal production fluids may change with time due to dilution and cooling of the reservoir fluids, resulting mainly from the invasion of cool and less-mineralized waters (Fig. 6). Regular measurement of the chemistry of liquid and vapor samples of fluids from selected wells provides information about changes in the reservoir. The data are also used to examine the need for and effectiveness of chemical dosing to prevent corrosion, and to monitor mineral deposition in the wells and pipework. However, interpretation of chemical changes is not necessarily straightforward because the chemistry of production fluids may vary between wells, and as the relative amounts of production from each well change with time (due to supply requirements) so the total chemistry may appear to change.



Geothermal Field and Reservoir Monitoring. Figure 3

Variation of temperature with depth in exploration drillholes in Tauhara geothermal field, New Zealand. Note the differences between wells and that in some wells (especially TH2, between -100 and -200 m) there are temperature reversals due to cool inflows

Samples of geothermal fluid may be taken either from within the wellbore (i.e., down-hole) or at the surface. Down-hole samples are taken using a special sampling device that captures the geothermal fluid in its in situ, undisturbed state; i.e., before phase separation, irreversible thermal cycles, or chemical reactions have occurred [10, 11]. Samples are usually taken at different depths in a well to determine the changes that occur as the fluid ascends the well. At the surface, a small sample of the two-phase geothermal fluid is drawn off from the flowing well and passed through a small separator which separates the liquid (water) from the vapor (steam). The liquid (water) fraction is then cooled by passing it through a



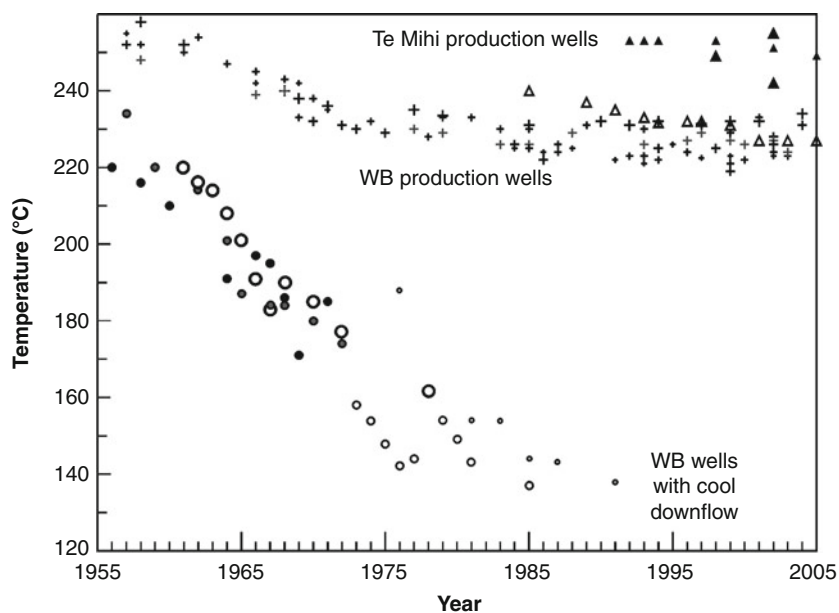
Geothermal Field and Reservoir Monitoring. Figure 4

Sketch showing variation of temperature with depth in some wells (Eastern borefield), at different times during early development of the Wairakei geothermal field, New Zealand (Taken from [2])

water-cooled coil and analyzed by standard techniques. The vapor (steam) fraction is passed into an evacuated glass flask containing a caustic solution to absorb acidic gases (carbon dioxide, hydrogen sulfide); the solution is then analyzed by titration. Trace gases (Ar, He, N) remain in the top of the flask and are removed and analyzed using a gas chromatograph [12].

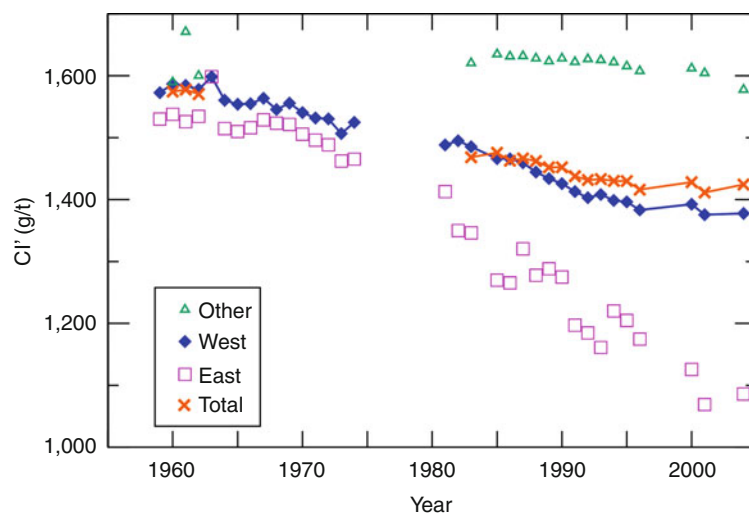
In high-temperature liquid-dominated geothermal systems, chloride is a major chemical species and an important indicator of changes in the reservoir fluid. A decrease in the chloride content may indicate dilution due to an influx of cold groundwater (Fig. 7), and an increase may indicate injection returns.

In fields where the geothermal fluid resides or passes through limestone, the fluid may become saturated in calcium carbonate (CaCO_3), which precipitates as calcite in the wellbores and pipelines as temperatures and pressures decline. This precipitate, known as scaling, is a serious problem because it



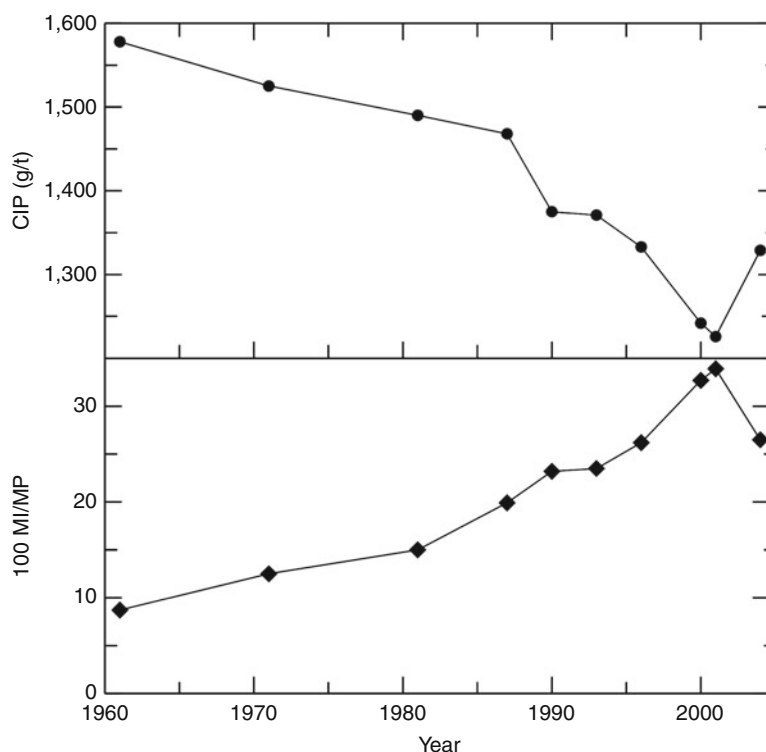
Geothermal Field and Reservoir Monitoring. Figure 5

Changes in production temperatures with time in Wairakei geothermal field (Taken from [3]). WB: Western Borefield. Note the large decrease in temperature of fluid from wells having cool downflows, whereas other production wells have experienced only a small temperature decline and temperatures have remained near constant since about 1975



Geothermal Field and Reservoir Monitoring. Figure 6

Changes with time in average chloride content (corrected for aquifer steam loss) of production fluid from different parts of Wairakei geothermal field (Taken from [9]). "West" and "East" refer to Western and Eastern borefields, respectively; "other" refers to the Te Mihi borefield, from where production began in the early 1980s



Geothermal Field and Reservoir Monitoring. Figure 7

Changes with time of the average chloride content of production well fluid (CIP) at Wairakei (upper) and in the amount of cold inflow (%) into the production zone computed from the chloride changes (lower) (Taken from [9]). The increase in the chloride value after 2000 is attributed to a larger proportion of deep water coming from a new production area (Te Mihi) and to more injected water (beginning 1995) reaching the production wells

reduces the fluid flow and particles of precipitate may flake off and enter the turbines causing damage. Monitoring of the carbonate helps identify problem wells and enable scaling rates to be determined.

Tracer Tests

Although primarily used for reservoir characterization purposes, tracer tests [13–15] may be used for reservoir management, particularly if major changes are made in the development of a field. A tracer test involves inserting a finite slug of a chemical or radioactive material (“tracer”) into an injection well and measuring the time for it to appear, and its concentration, in production and monitoring wells. Tracer tests to evaluate the flow patterns between injection and producing wells are common practice in oil and gas field operations. A wide variety of

chemical tracers have been used including: hydrofluorocarbons (tetrafluoroethane, trifluoromethane), naphthalene disulfonate, noble gases (neon, xenon), potassium halides (KBr, KI), rhodamine WT, and fluorescein. Fluorescein is the most commonly used tracer in liquid-dominated geothermal reservoirs because it is sufficiently stable to be used in reservoirs as hot as 250°C; it has a detection limit of approximately ten parts per trillion using conventional spectrofluorimetry; and can be detected using a simple, inexpensive, and easily operated filter fluorometer. Iodine 131 has been used as a radioactive tracer [16]. Some tracers travel preferentially in the vapor (steam) phase, others in the liquid (water) phase. By repeating tracer tests it may be possible to determine changes in fluid flow paths, particularly “short circuiting” of injected fluids from new injection wells directly to production wells.

Surface Monitoring

Flow Rate

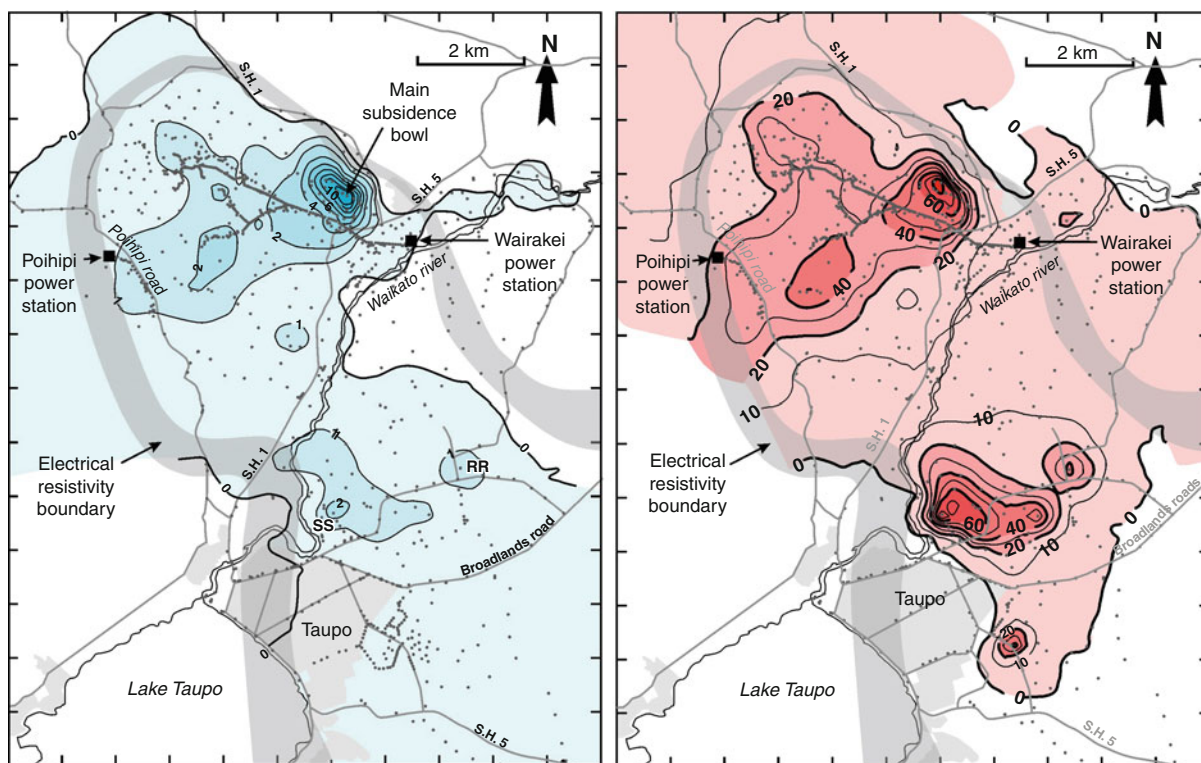
Monitoring fluid flow rates at the wellhead of production and injection wells is a basic monitoring tool which can indicate changes in the performance of individual wells and of total field performance. Sudden unexpected changes in flow rate from an individual well may indicate damage to the well casing. Gradual decreases in flow rate may indicate a fall in reservoir pressure in the vicinity of the well feed zone(s), or a change in the relative contributions of supply from different feed zones.

Ground Surface Movements

In a few geothermal fields, notably some of those in New Zealand, there have been significant ground

subsidence (up to 15 m) and horizontal deformation (up to 2 m) associated with production.

At Wairakei field (New Zealand), deformation was originally noticed when concrete drains became broken, and subsequently pipework has been mounted on roller supports to accommodate movement, although from time to time it has been necessary to remove and insert sections of pipe. Vertical deformation is measured by repeat surveys using an optical level to measure changes in elevation between permanent reference points such as benchmarks, referenced to a stable point outside the field. The frequency of surveys depends on the rate of subsidence and the location of the subsidence area. At Wairakei, the main steam lines are leveled every 2 years, and the whole field about every 4 years. In some fields, where there are not extensive amounts of surface vegetation, it has been possible to determine subsidence using interferometric synthetic



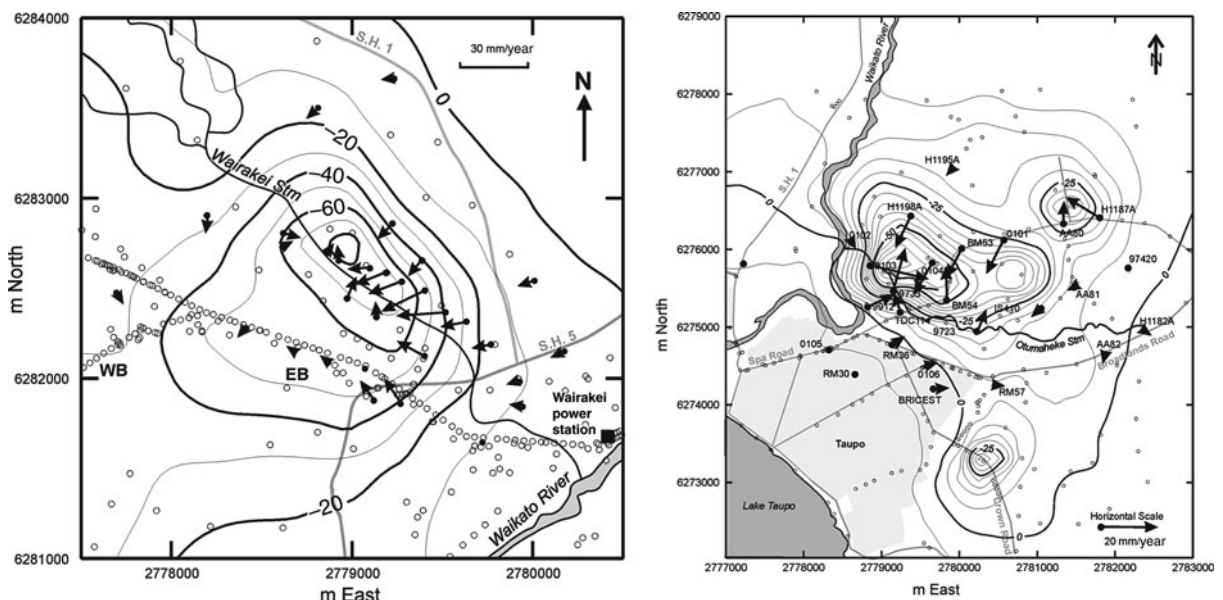
Geothermal Field and Reservoir Monitoring. Figure 8

Ground subsidence (left; m) and subsidence rates (right; mm/year for 2001–2005) at Wairakei-Tauhara geothermal field, New Zealand (Taken from [21]). Dots indicate survey points. Note that the intense subsidence is confined to several isolated subsidence bowls: Main; SS Spa Sights; RR Rakanui Road

aperture radar (InSAR) [17–20]. At Wairakei-Tauhara field (New Zealand) there have been small areas of intense subsidence within the field since measurements began shortly after production started in 1958 [21]. In this field the subsidence has occurred mainly in about four localized “subsidence bowls,” but over the remainder of the field the subsidence has been less than 1 m (Fig. 8). At the center of the main subsidence bowl the rate of subsidence increased to a maximum of over 450 mm/year in the late 1970s but has since decreased to about 50 mm/year (Fig. 8). The location of the bowls and their centers does not correspond to areas of maximum production; the center of the main subsidence bowl lies about 500 m from the original area of production. The cause of the subsidence at Wairakei-Tauhara and Ohaaki is associated with draining, and consequent compaction, of rocks of locally high compressibility within formations above the reservoir, due to a decrease in pressure within the steam zone in the upper part of the reservoir. Casing deformation indicates these rocks lie at about 100–300 m depth. However, the reason for the localized distribution of high compressibility in these rocks remains a puzzle; other

parts of the same formations do not have high compressibility. At Mokai field (New Zealand), there has been subsidence of up to 0.20 m around the injection wells, associated with cooling and thermal contraction of rocks in the injection aquifer.

Horizontal deformation is measured using theodolites or Geodimeters to measure changes in angles or distances between permanent reference points, or using global positioning system (GPS) techniques. Generally the reference points are permanent markers specifically installed for the purpose. At Ohaaki field (New Zealand), these consist of a concrete post made from a drainage pipe (approximately 600 mm diameter), mounted vertically, set in a concrete pad, and filled with concrete. A threaded pipe is set in the upper surface of the post to allow a theodolite, Geodimeter, or a target to be mounted on the post. At Wairakei-Tauhara field (New Zealand), the largest horizontal movement rates have been 25–30 mm/year and have occurred at the edges of the main subsidence bowl where the lateral changes in subsidence (tilt) have been the greatest [21]. The horizontal movement vectors generally point toward the center of the subsidence bowls (Fig. 9). The overall



Geothermal Field and Reservoir Monitoring. Figure 9

Horizontal deformation vectors (arrows) at subsidence bowls in the Wairakei-Tauhara geothermal field (Taken from [21]). Solid contours indicate rates of ground subsidence (mm/year) determined at benchmarks (open circles). Note the vectors point to the center of the subsidence bowl, and have greatest amplitude on the flanks of the bowl

pattern of horizontal movement has not changed greatly with time, but the rates of movement have declined as the subsidence rates have decreased.

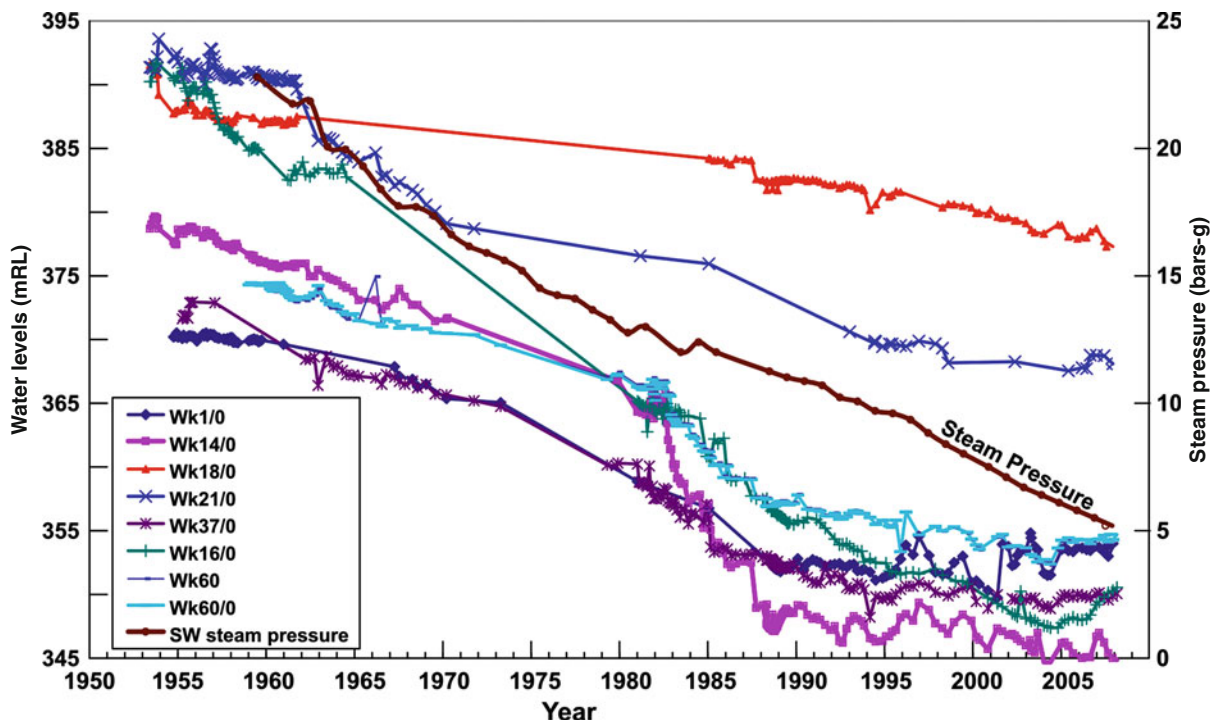
Groundwater

Near the ground surface above most geothermal reservoirs there is generally a complex sequence of groundwater aquifers containing cold or warm waters (and in places hot water and steam) which are often a source of potable water or used for industrial and domestic purposes (heating, cleaning). The aquifers are usually separated from each other by aquicludes.

These shallow aquifers can be affected by production from the deeper geothermal reservoir and hence many regulatory authorities require water levels, water chemistry, and temperatures in the aquifers to be monitored periodically. Monitoring is usually done using shallow wells drilled specifically for the purpose. These holes are generally about 3–5 cm diameter and are generally drilled vertically using a small truck-mounted auger. The holes usually have solid casing in the Vadose Zone and slotted

or screened casing from the water table to the bottom. Where several groundwater aquifers are present several monitor holes are drilled and care is taken in each to adopt a casing pattern that monitors a specific aquifer and ensures that the well does not result in interaction between separate aquifers, i.e., draining of an upper into a lower aquifer. In places where the ground temperature is less than about 50°C, plastic (PVC or ABS) casing is used, but for ground temperatures greater than this value steel casing is used. The open area of the screened casing should approximate the natural porosity of the rock formation, and the slots should widen inward to minimize plugging of the slots by fine formation material. Over a long period of time, fine silt and debris migrate through the screened casing and are deposited at the bottom of the hole; so the hole is generally drilled 5–10 m deeper than the natural water table.

Water levels are measured using a simple electric circuit device lowered down the well; this is powered by a small battery and contact with the water closes the circuit. Alternatively, a water level recorder can be installed which is comprised of a pressure transducer



Geothermal Field and Reservoir Monitoring. Figure 10

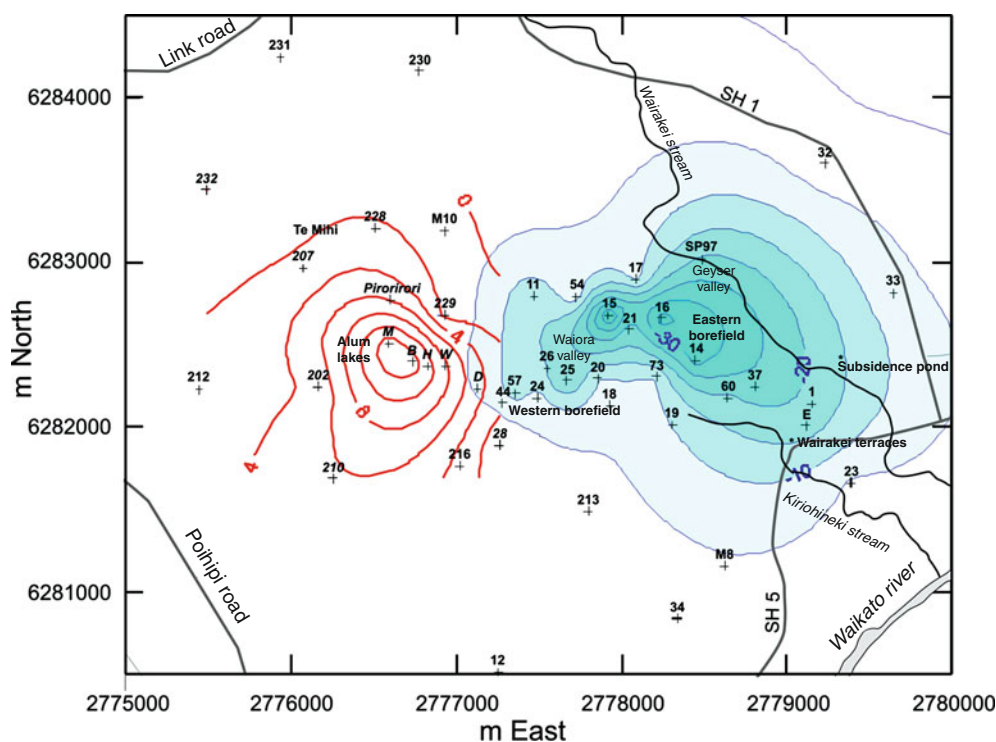
Changes in shallow groundwater level with time in the Eastern Borefield at Wairakei geothermal field (Taken from [22])

coupled to a data logger. Measurements are generally made at set times during the year to determine and correct for seasonal variations.

Changes in groundwater level (piezometric surface) can occur as a result of pressure declines in the deep geothermal reservoir. At Wairakei (New Zealand), decreases in groundwater level of up to 30 m have been recorded in the Eastern Borefield, an area where thermal features were fed by conduits from the deep reservoir [22] (Fig. 10). As pressures in the upper part of the reservoir decreased, the flow of geothermal fluid up conduits to the surface declined and eventually ceased. This allowed shallow groundwater to drain down the conduits, resulting in local regions of depression of the groundwater surface (Fig. 11). In some cases, where the near-surface geology is complex, these changes can be localized, especially where perched groundwater aquifers are present [23].

The temperature of groundwater is measured in shallow monitor holes using a digital thermometer and probe. Sometimes the temperature is measured not only at the water surface but also deeper in the monitor hole, to enable a temperature profile in the water to be obtained.

Samples for chemical analysis are obtained from groundwater monitor holes after water level and temperature measurements have been made. However, care must be taken not to sample stagnant water in these holes; only after five to ten wellbore volumes of water have been removed and naturally replaced should a sample be collected. Removal of stagnant water and collection of the samples is generally done using a small portable electric pump. Parameters that are usually measured are: pH, chloride, lithium, sodium, potassium, magnesium, sulfate (SO_4), total silica (SiO_2), total bicarbonate (HCO_3),



Geothermal Field and Reservoir Monitoring. Figure 11

Changes in shallow groundwater level (m) in at Wairakei geothermal field (Taken from [22]). Changes in the Eastern Borefield (blue contours) are for the period 1956–1995; changes in the Alum Lakes area (red contours) are for the period 1999–2006. Crosses indicate monitor wells; labels indicate well numbers (see Fig. 9); letters in the Alum Lakes area indicate thermal pools

and fluoride. In addition, measurements of stable isotopes $\delta_{18}\text{O}$, $\delta_2\text{H}$, and tritium are sometimes made.

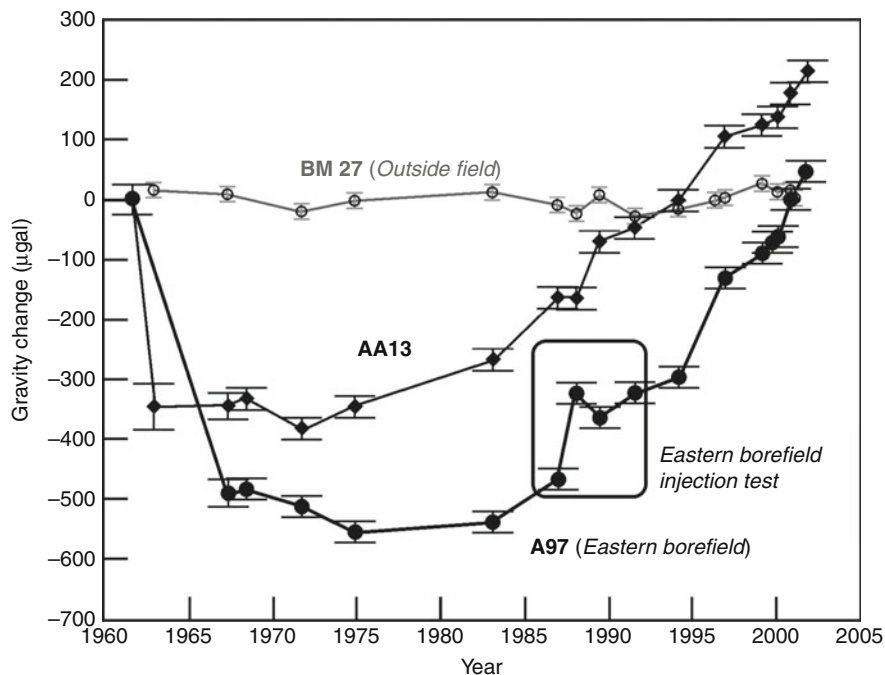
Microgravity

Exploitation of high-temperature geothermal resources usually involves withdrawing fluid from one area (production area) and, after using it to generate electricity or provide heat, injecting the liquid back into the ground in another area (injection area). This generally results in changes in mass (and corresponding density changes) in these areas, and hence small changes in the force of gravity at the surface (Figs. 12 and 13). The amount of gravity change in the production area will be related mainly to the amount of recharge, and to changes in the proportions of liquid water and steam in the production zone.

The changes in gravity are measured at permanent reference points such as survey benchmarks throughout and beyond the field boundaries using a portable

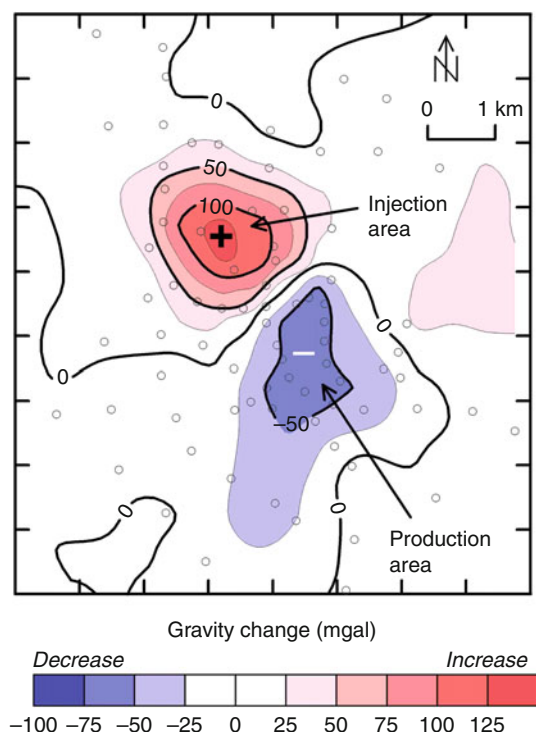
gravity meter. Generally a relative type of gravity meter is used which measures differences in gravity from a stable, reference point outside the geothermal field, although in some instances absolute gravity meters have been employed. Measurement precision is generally $5\text{--}10\ \mu\text{Gal}$ ($5\text{--}10 \times 10^{-8}\ \text{m/s}^2$). A baseline survey is made at 50–150 points prior to exploitation, and the survey repeated at intervals of 2–5 years afterward. Within each survey, corrections are made for the gravitational effects of changes in the position of the Moon and Sun (Earth tide) and for tares (jumps in zero point of the meter resulting from knocks). In determining the gravity differences between surveys (gravity changes) the data are corrected for the gravitational effects of ground elevation changes (subsidence), gravity changes at the reference point, and changes in ground-water level and temperature.

From the gravity changes it is possible to determine a field-wide value for recharge by numerical integration of the changes and application of Gauss's Theorem [25]. This method is completely independent of any



Geothermal Field and Reservoir Monitoring. Figure 12

Changes in gravity (μgal) associated with production at Wairakei geothermal field, New Zealand. Benchmarks A97 and AA13 lie in the Eastern borefield from which most production was obtained between 1958 and the 1980s. Note the increase in gravity since the late 1970s, associated with resaturation of the production aquifer (Taken from [24])



Geothermal Field and Reservoir Monitoring. Figure 13 Gravity changes at Mokai geothermal field, New Zealand, between 1997 and 2004. Open circles indicate measurement points; contour interval $25 \mu\text{gal}$ ($25 \times 10^{-8} \text{ m/s}^2$). Data has been corrected for the effects of small amounts of ground subsidence

assumptions about fluid density, depth of production, permeability, or porosity; its accuracy is limited only by the precision of the gravity measurements, and errors inherent in the integration of the data. Gravity change measurements also may provide: field-wide and local values for recharge of fluid into a geothermal system; information about changes in saturation in different parts of the two-phase zone; a test of complex, three-dimensional, numerical reservoir simulation models for exploitation of a field [26, 27]; information about the location and movement of injected fluid [28]; and estimates of reservoir parameters such as permeability (k), permeability-thickness (kh), and storativity (ϕ_{ch}) [29].

Gravity change data can also be used to discriminate between two (or more) numerical reservoir simulation models for exploitation of a field. Such models are important in guiding development of a field. The models

for high-temperature liquid-dominated fields predict the development and extension of 2-phase conditions and subsequent changes in saturation (and hence density and gravity changes) which involve assumptions about the geometry of the field, various reservoir properties, and behavior of the field during exploitation. Discrepancies between the theoretical (model-derived) and measured gravity changes may indicate that assumptions made in setting up the models are wrong.

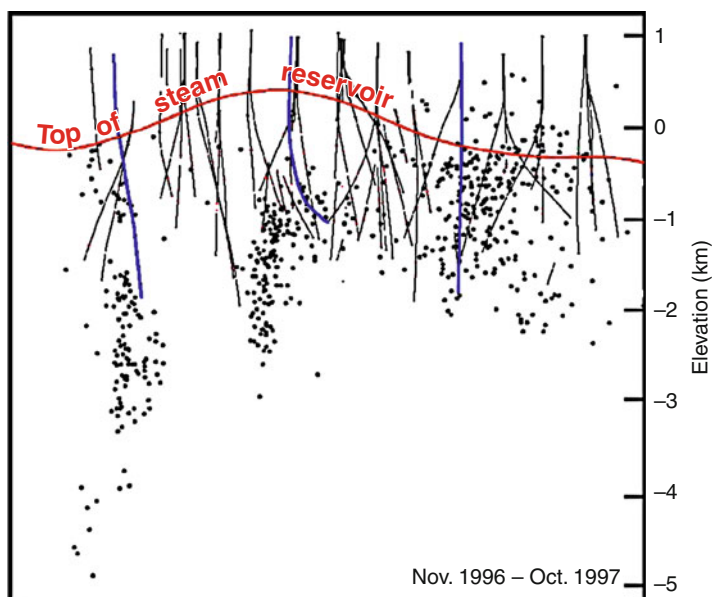
Another important use of gravity change data is to track the path of injected water. If the waste liquid water is injected into a region of 2-phase conditions the liquid is cooler, and hence denser than the fluids present, and tends to sink toward the bottom of the zone. If the rocks do not have isotropic permeability the liquid will move more rapidly along paths of high permeability and in response to any pressure gradients that might be present; this movement being reflected by the gravity changes.

Electrical Resistivity

Most high-temperature geothermal fields, particularly liquid-dominated fields, are delineated by a boundary zone in which the electrical resistivity increases from low values ($1\text{--}50 \Omega\text{m}$) inside the field to high values ($>200 \Omega\text{m}$) outside the field. If production from the field causes significant decreases in fluid pressure in the upper part of the reservoir, there may be an influx of cool water from outside the field. If such an influx is large enough and at shallow depth then there may be an apparent lateral shift in the electrical resistivity boundary zone where this influx is occurring. Such a shift may be detected, before the cool water reaches production wells, by repeating electrical resistivity surveys across the boundary zone. Another situation where repeating electrical resistivity surveys may be useful is where hot saline waste water is injected into cold water aquifers outside the field.

Induced Seismicity

In many high-temperature geothermal fields exploitation can result in an increase (above the normal background) in the number of small magnitude earthquakes (micro-earthquakes) within the field (Fig. 14) [30]. Induced seismicity occurs in both liquid- and



Geothermal Field and Reservoir Monitoring. Figure 14

Cross section through The Geysers field (California, USA) showing locations of earthquakes (*black dots*) during a 12-month period. Injection wells are shown in blue. Earthquake hypocenters and wells within 2,000 ft (600 m) of the section line have been projected onto the cross section. Note that the earthquake hypocenters extend to depths greatly below the bottom of some injection wells (Taken from [34])

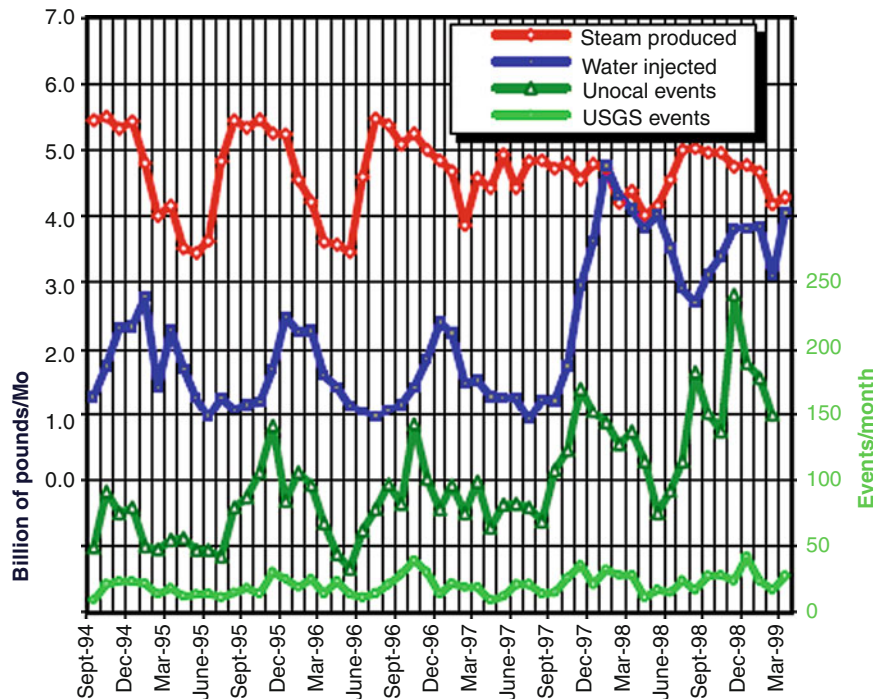
vapor-dominated high-temperature fields, and in enhanced geothermal systems, but has rarely been observed in low-temperature fields. The increase is caused mainly by injection because when injection starts or is increased the number of local micro-earthquakes increases, and when injection decreases or is stopped the number of small earthquakes decreases [30, 31]. The main cause of the micro-earthquakes is high wellhead injection pressures that increase the pore pressure at depth, particularly in existing fractures, which allows movement to suddenly release stress and generate an earthquake. Thermal stress associated with the injection of cool waste water into a hot fluid aquifer may also trigger earthquakes in the vicinity of the injection wells.

The micro-seismicity is generally monitored by an array of seismometers (vertical- or three-component) placed in shallow drillholes (to minimize the effects of anthropogenic “noise”). Usually the signals from each seismometer are telemetered in real time to a central recording apparatus – this ensures consistent relative timing of the signals which is critical for determining

the location (hypocenter and epicenter) of the earthquake. The hypocenter of each seismic event is then computed from the relative time differences of the arrival of the shock wave at each seismometer, assuming a specific local seismic velocity model [32]. The seismic velocity model is calculated either by inverting the seismic data collected over a period of time [33], or from explosions set off in drillholes. The magnitudes of the micro-earthquakes within the geothermal field are determined by comparison with the magnitudes of large local events as determined by national or regional seismic networks.

During a 4½ year period in which the mass of water injected at The Geysers field increased and decreased (due to seasonal power loading), the number of events measured by a detailed seismic network (Fig. 15) appeared to be related more closely to the injection rates rather than (steam) production rates [34].

Real-time monitoring of micro-seismicity can be used to minimize the felt intensity of shaking by



Geothermal Field and Reservoir Monitoring. Figure 15

Comparison of changes in the number of seismic events detected with variation in the amount of injection at The Geysers field, USA (Graph taken from [34])

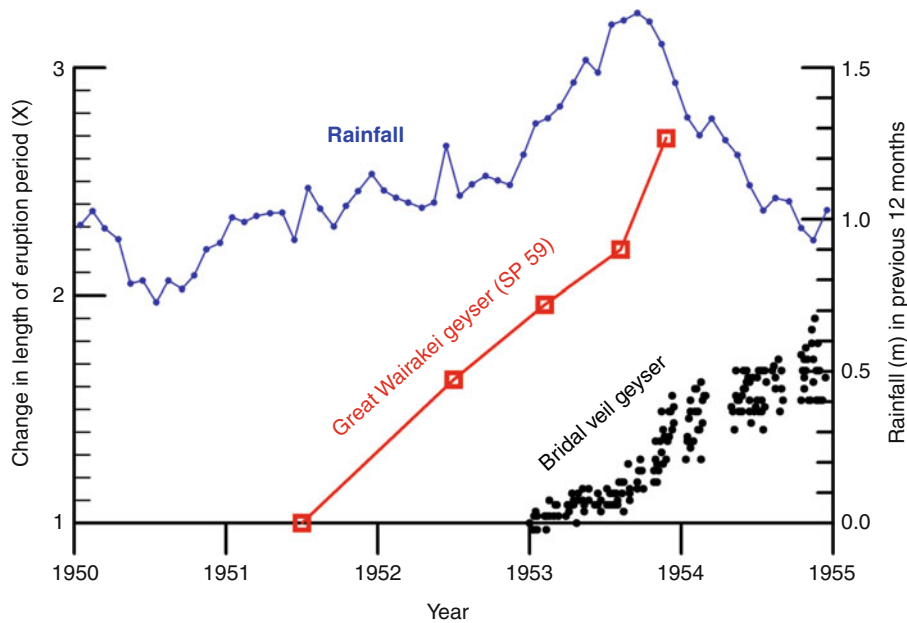
reducing or stopping injection if the number of events exceeds certain predetermined thresholds – this is known as “traffic light control” of injection [30, 35].

Thermal Features

Many geothermal fields, especially high-temperature liquid-dominated fields, are manifested at the surface by natural thermal features such as geysers, hot pools, hot springs, mud pools, fumaroles, and areas of thermal ground. Changes in these features with time during production from the field can indicate changes in the geothermal reservoir from which fluid is being withdrawn, although in some cases it may be difficult to separate natural changes from production-induced changes. A wide variety of thermal features can be monitored, however, the most sensitive and easiest to monitor are geysers, hot pools, and hot springs.

Geysers Geysers occur in high-temperature geothermal fields. They are the most spectacular and the most valued of natural thermal features (for cultural and economic reasons) and are the most sensitive to production-induced changes in a geothermal system. Geysers are generally monitored by measuring changes in the eruption period (time between the start of successive eruptions), usually by a simple device that continuously measures the temperature of water in a channel leading from the geyser. Increases in the eruption period of geysers may be indicative of pressure decreases in the reservoir: increases in the eruption period of two geysers at Wairakei field were measured (Fig. 16) prior to their demise during the time of preproduction well testing (test discharge period) [36].

It is difficult to measure the volume of erupted water because of flashing to steam during the eruption, and evaporation from or absorption into the rocks surrounding the vent after the erupted water has fallen to the ground. If the geyser erupts frequently, it is



Geothermal Field and Reservoir Monitoring. Figure 16

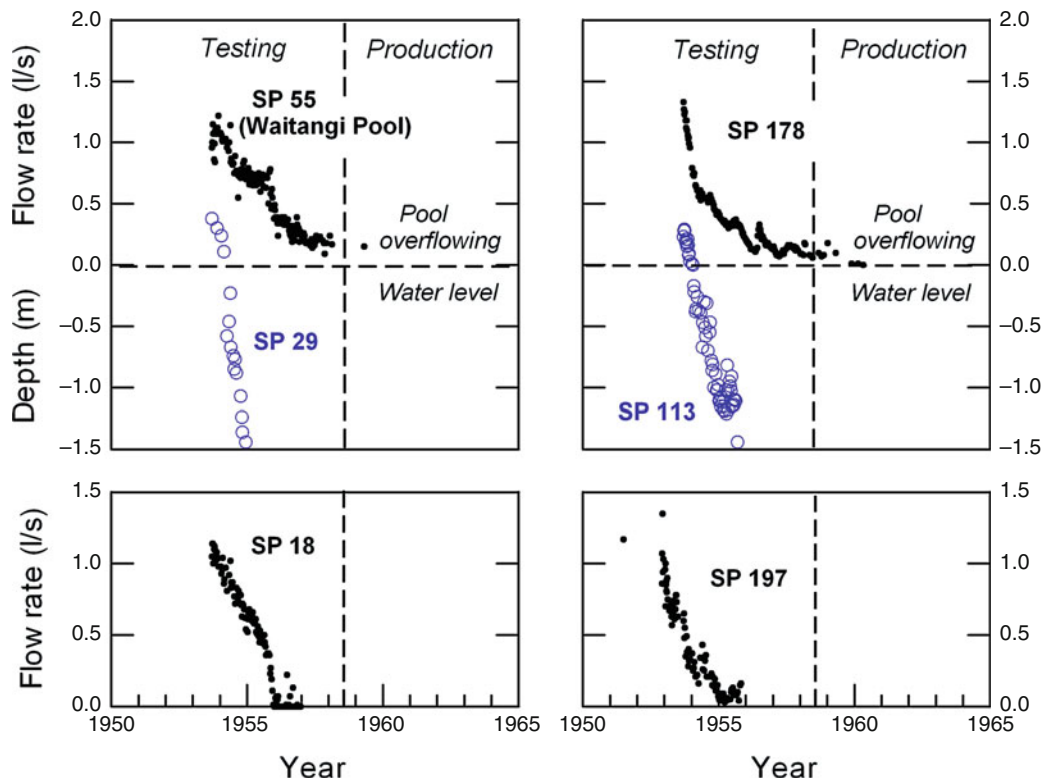
Changes in length of eruption period (T/T_0) of two geysers in Geyser Valley at Wairakei geothermal field (New Zealand) during the time of preproduction test discharges. Eruption periods are normalized to $T_0 = 12.5$ h for the Great Wairakei Geyser, and $T_0 = 39$ min for the Bridal Veil Geyser. Rainfall data are monthly running totals of rainfall in the previous 12 months. Note the steady increase in length of eruption period with time (Taken from [36])

feasible to measure eruption height by using a video camera; however, the volume and height of geyser eruptions often vary naturally due to wind gusts and to seasonal changes in rainfall.

Springs and Hot Pools Hot springs and hot pools are also associated mainly with high-temperature geothermal fields and have important cultural and economic value [37]. Regular or continuous monitoring of the temperature, chemistry and flow rate of hot springs, and the temperature and water level in hot pools are generally made. Decreases in these parameters are also indicative of changes which may lead to the demise of these features.

The temperature of water emerging from the ground in hot springs and in pools is measured using a variety of commercially available devices employing thermistors or thermocouples. The temperature data are often measured continuously and captured in a data logger. The flow rate of springs and the rate of outflow from hot pools are generally measured by constructing a channel to take all the water from the

feature and pass it through a V-notch weir. The basic principle is that flow rate is directly related to the water depth above the bottom of the V. The V-notch design causes small changes in flow rate to have a large change in depth allowing more accurate measurement than with a rectangular weir. From the measurement of the height of the water flowing through the V-notch and the angle of the V, the flow rate can be calculated. However, the value obtained may need to be adjusted to take into account rainfall and evaporation from the surface of the pool and channel before the water reaches the V-notch. To monitor changes in the chemistry of the water, samples are taken and analyzed in a laboratory; usually chloride content is the main chemical species measured. If flow into a hot pool decreases sufficiently such that evaporation exceeds inflow, then the water level in the pool may fall below the overflow and the water level may temporarily or permanently fall. A difficulty in interpreting temperature and flow rate data is separating natural changes from production-induced changes; this can be minimized by taking the measurements at a frequency



Geothermal Field and Reservoir Monitoring. Figure 17

Changes in outflow rate and water level with time in some hot pools at Geyser Valley, Wairakei (Taken from [36]). Note how, as well testing proceeded, the outflow rates declined and the springs stopped flowing. In the pools associated with springs SP 29 and SP 113, the water level dropped below the outlet until water stopped flowing and they dried up

sufficient to determine natural changes caused by changes in rainfall and groundwater level.

At Wairakei geothermal field, the flow rate from hot springs in Geyser Valley declined (Fig. 17) during the time of preproduction test discharges of exploration wells in the Waiora Valley [36]. Initially, the changes were small and isolated and were thought to be caused by natural climatic variations. It was not until much later that it was recognized that the changes to the hot springs were associated with changes in the deep reservoir resulting from fluid withdrawal some distance away. A decline in thermal features in producing high-temperature geothermal fields appears to be associated mainly with a decline in reservoir pressure. As the pressure declines, so also does the amount of geothermal fluid reaching the surface and hence the thermal features decline in size and vigor. If pressures fall further then the features may die and the flow may

reverse, with cold groundwater flowing down into the reservoir; once this situation has occurred it may take a long time to resurrect the features. Monitoring of changes to hot springs and hot pools may enable declines to be recognized quickly and remedial action taken.

Future Directions

Reservoir monitoring will probably expand in scope and increase in frequency in the future because regulatory authorities are generally becoming more concerned about environmental effects. There are also commercial and economic effects which may result in more monitoring. Monitoring data help improve numerical simulation models which developers use to identify the potential effects of changes in production and injection. The modeling is not only for planning and operational

purposes but also to help secure loan funding for future expansion at the most favorable interest rates because bankers seek to reduce risk, and monitoring and modeling help in the reduction of risk.

Significant improvements in monitoring are likely to be the development of down-hole instrumentation capable of withstanding high temperatures for long periods of time. However, the problem of relating what is measured in a drillhole to what is occurring in the rock outside the hole will still remain.

Bibliography

Primary Literature

1. Thain IA, Carey BS (2009) Fifty years of geothermal power generation at Wairakei. *Geothermics* 38(1):48–63
2. Allis RG, Hunt TM (1986) Analysis of exploitation induced gravity changes at Wairakei geothermal field. *Geophysics* 51:1647–1660
3. Bixley PF, Clotworthy AW, Mannington WI (2009) Evolution of the Wairakei geothermal reservoir during 50 years of production. *Geothermics* 38(1):145–154
4. Yoder JL (1998) Using meters to measure steam flow. http://www.flowresearch.com/articles/Plant_Engineering_0498.pdf
5. Mattar WM (2005) Advances in Coriolis technology resolve tough pipeline flow measurement challenges. *Pipeline Gas J*, July 2005. <http://www.pipelineandgasjournal.com>
6. Bixley PF, Wilson DM (1985) Rapid casing corrosion in high temperature liquid dominated geothermal fields. In: Proceedings of the 10th Workshop on Geothermal Reservoir Engineering, Stanford, pp 35–40
7. Bowyer D, Bignall G, Hunt T (2008) Formation and neutralization of corrosive fluids in the shallow injection aquifer, Rotokawa geothermal field, New Zealand. *GRC Trans* 32:201–205
8. Bixley PF, Hattersley SD (1983) Long term casing performance of Wairakei production wells. In: Proceedings of the 5th NZ Geothermal Workshop, Auckland, pp 257–263
9. Glover RB, Mroczek EK (2009) Chemical changes in natural features and well discharges in response to production at Wairakei, New Zealand. *Geothermics* 38(1):117–133
10. Klyen LE (1973) A vessel for collecting subsurface water samples from geothermal drillholes. *Geothermics* 2:57–60
11. Arnórsson S, Bjarnason JÖ, Giroud N, Gunnarsson I, Stefánsson A (2006) Sampling and analysis of geothermal fluids. *Geofluids* 6(3):203–216
12. Grob RL, Barry EF (2004) Modern practice of gas chromatography, 4th edn. Wiley-Interscience, New York, 1064 pp. ISBN-10: 0471229830, ISBN-13: 978-0471229834
13. Rose PE, Apperson KD, Johnson SD, Adams MC (1997) Numerical simulation of a tracer test at Dixie Valley, Nevada. In: Proceedings of the 22nd Workshop on Geothermal Reservoir Engineering, Stanford, pp 169–176
14. Adams MC, Beall JJ, Enezy SL, Hirtz PN, Kilbourn P, Koenig BA, Kunzman R, Smith JL (2001) Hydrofluorocarbons as geothermal vapor-phase tracers. *Geothermics* 30(6):747–775
15. Hirtz PN, Kunzman RJ, Broaddus MK, Barbitta JA (2001) Developments in tracer flow testing for geothermal production engineering. *Geothermics* 30(6):727–745
16. McCabe WJ, Barry BJ, Manning MR (1983) Radioactive tracers in geothermal underground water flow studies. *Geothermics* 12:83–110
17. Massonnet D, Holzer T, Vadon H (1997) Land subsidence caused by the East Mesa geothermal field, California, observed using SAR interferometry. *Geophys Res Lett* 24:901–904
18. Massonnet D, Feigl KL (1998) Radar interferometry and its application to changes in the earth's surface. *Rev Geophys* 36:441–500
19. Hsing-Chung C, Linlin G, Rizos C (2005) InSAR and mathematical modelling for measuring surface deformation due to geothermal water extraction in New Zealand. In: Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium, vol 3, pp 1587–1589
20. Hole JK, Bromley CJ, Stevens NF, Wadge G (2007) Subsidence in the geothermal fields of the Taupo Volcanic Zone, New Zealand from 1996 to 2005 measured by InSAR. *J Volcanol Geoth Res* 166:125–146
21. Allis RG, Bromley CJ, Currie S (2009) Update on subsidence at the Wairakei-Tauhara geothermal system. *Geothermics* 38(1):169–180
22. Bromley CJ (2009) Groundwater changes in the Wairakei-Tauhara geothermal system. *Geothermics* 38(1):134–144
23. Bromley CJ, Hunt TM, Morris C (1993) Cold downflows of groundwater at Ohaaki Geothermal Field; preliminary results. In: Proceedings of the 15th New Zealand Geothermal Workshop, Auckland, pp 181–186
24. Hunt TM, Bromley CJ, Risk GF, Sherburn S, Soengkono S (2009) Geophysical investigations of the Wairakei Field. *Geothermics* 38(1):85–97
25. Hunt TM (1970) Gravity changes at Wairakei Geothermal Field, New Zealand. *Geol Soc Am Bull* 81:529–536
26. Atkinson PG, Pedersen JR (1988) Using precision gravity data in geothermal reservoir engineering modelling studies. In: Proceedings of the 13th Workshop on Geothermal Reservoir Engineering, Stanford, pp 35–40
27. Hunt TM, Allis RG, Blakely MR, O'Sullivan MJ (1990) Testing reservoir simulation models for the Broadlands Geothermal Field using precision gravity data. *Geoth Resour Counc Trans* 14:1287–1294
28. Hunt TM (2005) Using repeat microgravity measurements to track reinjection in a liquid-dominated field. In: Proceedings of the World Geothermal Congress 2005 (CD). Paper No. 1117
29. Hunt TM, Kissling WM (1994) Determination of reservoir properties at Wairakei Geothermal Field using gravity change measurements. *J Volcanol Geoth Res* 63:129–143
30. Majer EL, Baria R, Stark M, Oates S, Bommer J, Smith B, Asanuma H (2007) Induced seismicity associated with Enhanced Geothermal Systems. *Geothermics* 36:185–222

31. Sherburn S, Allis RG, Clotworthy A (1990) Microseismic activity at Wairakei and Ohaaki geothermal fields. In: Proceedings of the 12th NZ Geothermal Workshop, Auckland, pp 51–55
32. Stein S, Wyssession M (2002) An Introduction to seismology, earthquakes, and earth structure. Wiley-Blackwell, Oxford, 512 pp. ISBN: 978-0-86542-078-6
33. Zucca JJ, Hutchings LJ, Kasameyera PW (1994) Seismic velocity and attenuation structure of the Geysers geothermal field, California. *Geothermics* 23(2):111–126
34. Bommer JJ, Oates S, Cepeda JM, Lindholm C, Bird J, Torres R, Marroquin G, Rivas J (2006) Control of hazard due to seismicity induced by a hot fractured rock geothermal project. *Eng Geol* 83:287–306
35. Smith B, Beall J, Stark M (2000) Induced seismicity in the SE Geysers field, California, USA. In: Proceedings of the World Geothermal Congress 2000, Kyushu-Tohoku, Japan, pp 2887–2892
36. White PA, Hunt TM (1996) Simple modelling of the effects of exploitation on hot springs, Geyser Valley, Wairakei, New Zealand. *Geothermics* 34:184–204
37. Cataldi R, Hodgson SF, Lund JW (1999) Stories from a heated Earth. Geothermal Resources Council and the International Geothermal Association, Sacramento, California, 569 pp. ISBN 0-934412-19-7

Books and Reviews

- Armstead HCH (1980) Geothermal energy: its past, present and future contributions to the energy needs of man. E & F.N. Spon, London
- Dickson MH, Fanelli M (2003) Geothermal energy utilization and technology. UNESCO Publishing, Paris, 205 pp. ISBN 92-3-103915-6
- DiPippo R (2008) Geothermal power plants: principles, applications, case studies and environmental impact, 2nd edn. Butterworth-Heinemann, Oxford, 520 pp. ISBN: 978-0-7506-8620-4
- Ellis AJ, Mahon WAJ (1977) Chemistry and geothermal systems. Academic, New York, 392 pp. ISBN: 0-12-237450-9
- Grant MA, Donaldson IG, Bixley PF (1982) Geothermal reservoir engineering. Academic, New York, 369 pp. ISBN: 0-12-295620-6
- Hunt TM (2001) Five lectures on environmental effects of geothermal utilization. Report 2000-1, United Nations University, Reykjavik, Iceland, 109 pp. ISBN-9979-68-070-9
- <http://os.is/Apps/WebObjects/Orkustofnun.woa/swdocument/2056/Trevor03.pdf>
- Kruger P, Otte C (1973) Geothermal energy: resources, production, stimulation. Stanford University Press, California, 360 pp
- Rybach L, Muffler LJP (1981) Geothermal systems: principles and case histories. Wiley, New York
- Thorhallsson S (2003) Geothermal well operation and maintenance. IGC2003 – Short Course September 2003:195–217. <http://unugtp.is/Apps/WebObjects/Orkustofnun.woa/swdocument/539/13Sverrir.pdf>

Geothermal Power Capacity, Sustainability and Renewability of

SUBIR K. SANYAL

GeothermEx, Inc., Richmond, CA, USA

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Concepts of Sustainability and Renewability

Relation Between Renewable and Sustainable Capacities

Estimation of Renewable, Sustainable and Commercial Capacities

An Illustrated Case History

Future Directions

Bibliography

Glossary

Discharge A measure of the flow rate of steam, water, or heat discharged at or near the ground surface from a subsurface geothermal reservoir.

Geothermal The naturally occurring heat found beneath the surface of the Earth, ultimately originating from the internal magmatic processes of the Earth's core. A geothermal energy project utilizes the hot water or steam found within certain large bodies of rock, referred to as a geothermal reservoir.

Power capacity The amount of energy produced per unit time, or the amount of the electric power capacity that a power generation facility is designed to produce.

Recharge Natural influx of hot fluids into a geothermal system.

Renewable A natural energy resource that is inexhaustible or can replenish itself over time.

Specific heat The amount of heat required, in calories, to raise the temperature of 1 g of a substance by 1°C.

Sustainable A natural energy resource which, if managed carefully, will provide the needs of a community or society indefinitely, without depriving future generations of their needs.

Thermal anomaly A departure from the normal or expected temperature in the subsurface as distinguished by geological, geophysical or geochemical means, which is different from the general surroundings.

Thermal conductivity A measure of the ability of a material to conduct heat.

Definition of the Subject and Its Importance

Geothermal energy is the heat energy of the earth, produced through wells as hot water or steam. Geothermal power capacity is this energy extraction rate (whether as thermal energy or equivalent electrical energy produced per unit time), expressed in Watt or an equivalent unit. The vast content of heat energy within the earth is limitless for all practical purposes, but the geothermal power capacity available from the earth is constrained by various technological and economic limits to the utilization of this energy. Given a geothermal power generation scheme (for example, a district heating scheme using geothermal water or an electric power operation using geothermal water or steam), the issue is how sustainable, technically and economically, the scheme would be and to what extent this energy supply is naturally renewable.

For the purposes of this entry, sustainability is defined as the ability to economically maintain an installed power capacity, over the amortized life of a power plant, by taking practical steps, such as, drilling “make-up” wells as needed to compensate for resource degradation (pressure decline, well productivity decline, or cooling of the produced hot water or steam). Renewability is defined here as the ability to maintain an installed power capacity indefinitely without encountering any resource degradation; this renewable power capacity at a geothermal site is generally too small for commercial development of electrical power capacity, but may be adequate for district heating or other direct uses of the geothermal energy.

Introduction

As per above definitions, it is argued below that only a portion of the sustainable geothermal power capacity at a site is renewable; yet, geothermal energy is widely believed to be entirely renewable. Therefore, it is important to objectively review the concepts of

sustainability and renewability of geothermal power capacity and their quantification. These issues are addressed below.

Concepts of Sustainability and Renewability

Many articles have been published on the renewability and sustainability of geothermal energy [1–7]. However, no universally accepted definitions of the words “renewability” and “sustainability” as regards geothermal power seem to exist and definitions used often have ambiguities. For example, Axelsson et al. [3] defines “renewable” generation capacity (Fig. 1) as:

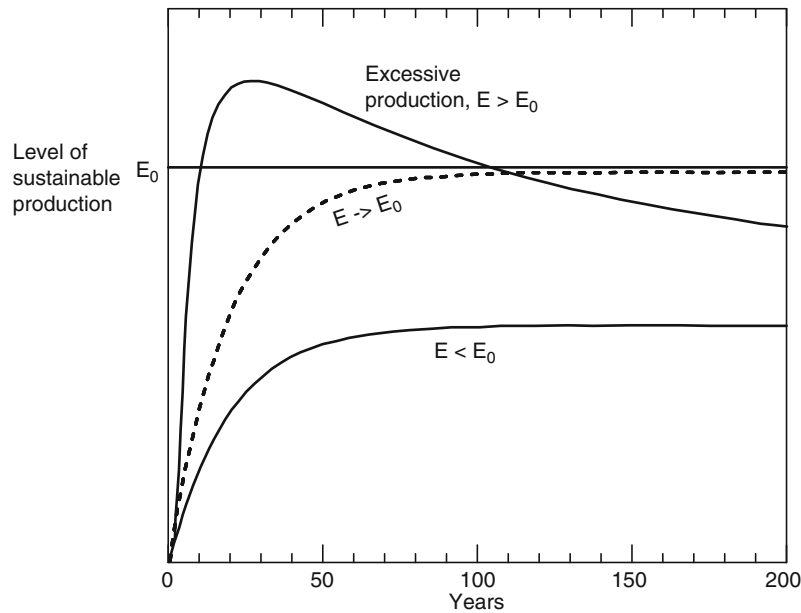
- The energy extracted from a renewable energy source is always replaced in a natural way by an additional amount of energy, and the replacement takes place on a similar time scale as that of the extraction.

And Axelsson et al. [3] define “sustainable” generation capacity as follows (Fig. 1):

- For each geothermal system, and for each mode of production, there exists a certain level of maximum energy production, E_0 , below which it will be possible to maintain constant energy production from the system for a very long time (100–300) years. If the production rate is greater than E_0 it cannot be maintained for this length of time. Geothermal energy production below, or equal to E_0 is termed **sustainable production**, while production greater than E_0 is termed **excessive production**.

An objective review of the above definitions is presented below, extracted largely from an earlier paper by the author.

The above definition of renewability essentially equates renewable capacity to the natural heat recharge rate (conductive plus convective) into a geothermal reservoir, which remains constant over geologic time (that is, tens of thousands of years) in the natural state. This recharge rate can be estimated for an actual reservoir by numerical simulation of the natural, steady-state heat flow, and measured temperature and pressure distributions, within the system. The renewable capacity is, however, frequently too small for commercial development because of the unfavorable economy of scale in capital and operation costs and relatively high cost of infrastructure development



Geothermal Power Capacity, Sustainability and Renewability of. Figure 1
Illustration of the definition of sustainable and excessive production levels [3]

associated with a small power project. The above definition of sustainability may perhaps be acceptable for non-electrical uses of geothermal energy (such as district heating), which are of relatively low intensity and are not capital-intensive, but the definition has inherent ambiguities and limitations for practical applications to the electric power industry. The difference between renewability and sustainability as defined above is a matter primarily of the time scale; as discussed later in connection with a case history presented below, an exploitation level that can be sustained for 100–300 years can most likely be sustained indefinitely. Therefore, for most fields, the above two definitions are essentially identical.

A constant energy production rate over a time span of 100–300 years is reasonable for defining renewability but not sustainability. A power plant can be sustained over a typical amortized life of 20–30 years at a capacity level much higher than the renewable capacity level by make-up well drilling or taking other steps to mitigate resource degradation. Numerical simulation consistently shows that any resource degradation caused over a typical plant life of 20–30 years would essentially disappear within a 100–300-year time frame after the project is shut down; the pressure would return to the original

level in about 20–30 years and the temperature within 100–300 years, the actual time taken being dependent on the natural convective heat recharge rate at the site (see, for example, [8]). Therefore, over a 100–300-year time span, commercial exploitation for 20–30 years at the sustainable level should not leave any permanent impact on the resource base. On the other hand, it is likely that producing the reservoir at a level higher than the renewable capacity estimated from natural-state modeling would actually cause an increase in the natural recharge rate of hot water into the reservoir. This has frequently been the author's experience from monitoring many producing geothermal fields; the case history discussed later illustrates this point. Therefore, estimate of renewable or sustainable power capacity from the simulation of the natural state of a geothermal reservoir is conservative; substantial production history is needed to estimate these capacities with any confidence. On the other hand, unless these capacities can be determined to the satisfaction of financial institutions, it is not possible to obtain long-term financing for a power plant; unless a power plant is installed, accumulation of substantial production history is out of the question. This is a fundamental conundrum of the geothermal power industry.

Geothermal reserves are normally expressed in terms of the installed capacity sustainable for the life of a power plant; empirical experience shows this reserve level to be an order of magnitude higher than the renewable level estimated from the natural state of

the reservoir; see, for example, [Table 1](#) (to be discussed later). Therefore, if the definition of sustainability in [Fig. 1](#), which is essentially same as renewability, is to be used, the geothermal resource base worldwide should be considered an order of magnitude smaller than is

Geothermal Power Capacity, Sustainability and Renewability of. Table 1 Empirical data on renewable and sustainable capacities [7]

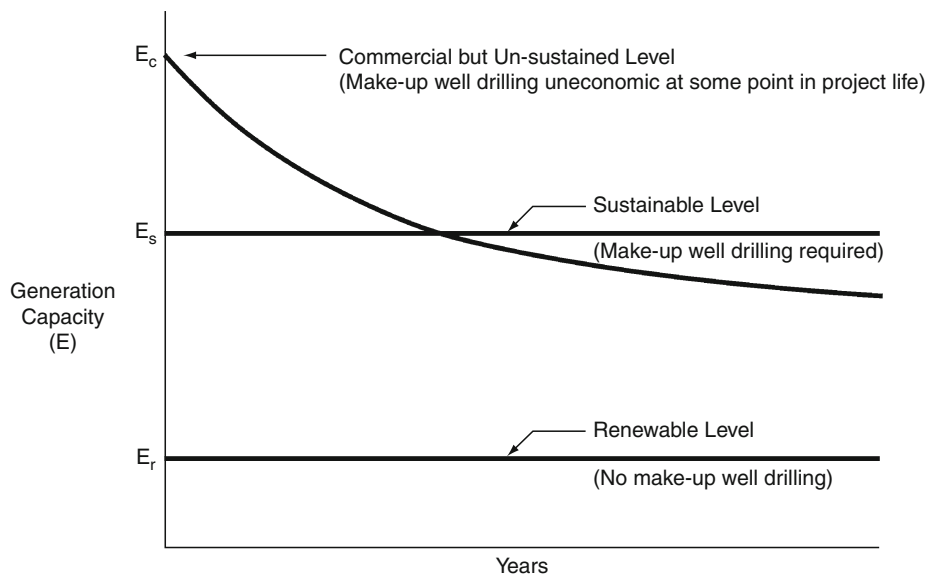
Field	Location	Renewable capacity (MWe)	Sustainable capacity (MWe)	References
Ahuachapan	El Salvador	24.8	95+	[9]
Beowawe	Nevada	1.3	13+	[10]
Cerro Prieto	Mexico	73.3	720	[11]
Desert Peak	Nevada	14	90+	[12]
Heber	California	1.7	70	[13]
Kakkonda	Japan	26.6	80+	[14]
Kawareu	New Zealand	15.5	230	[15]
Krafla	Iceland	5.3	60	[16]
Mammoth	California	25	90+	[17]
Mindanao	Philippines	9.6	102	[18]
Miravalles	Costa Rica	16.5	168	[19]
Mori	Japan	5.4	50	[20]
Mutnovsky	Russia	9.2	100	[21]
Nesjavellir	Iceland	16.6	160	[22]
Ngawha	New Zealand	2.5	30	[23]
Oguni	Japan	8.2	20+	[24]
Onikobe	Japan	2	25	[25]
Roosevelt Hot Springs	Utah	5.3	50+	[26]
San Emidio	Nevada	1.9	10+	[12]
Sibayak	Indonesia	11	30+	[27]
Soda Lake	Nevada	1.6	15	[12]
Stillwater	Nevada	4	40	[12]
Sumikawa	Japan	4	50+	[28]
Takigami	Japan	3	25	[29]
Uenotai	Japan	2.5	25	[30]
Wairakei	New Zealand	46	220+	[31]
Wasabizawa	Japan	5.6	40+	[32]
Zunil	Guatemala	2.44	25	[33]
		Total: 386	Total: 2,056+	

generally accepted today. In other words, exploitation of geothermal resources would be artificially constrained to an order of magnitude lower than the level at which exploitation is readily possible without any long-term negative impact on the resource base. This would make development of many fields for power generation economically prohibitive. Furthermore, this cannot be a socially responsible position considering that a higher rate of exploitation can only reduce the current fossil fuel usage, thus reducing environmental pollution today and saving fossil fuel resources for future generations. As discussed below, there is social virtue in preserving more of fossil fuel resources for the future, and instead, maximizing the use of power from geothermal resources, which are renewable within the 100–300-year time frame.

While geothermal power has far less environmental impact than power from fossil fuels, it is inevitable that power derived from fossil fuels will become progressively more environmentally benign in the future. Finally, unlike geothermal, fossil fuels also serve as raw material for petrochemicals and coal-based organic chemicals. While future generations may harness hitherto unforeseen sources of energy, fossil fuels will still be needed as raw material for chemicals. Therefore, one can justify a higher rate of geothermal power use

today than adhering to a level that is renewable within the 20–30-year lifetime of a power plant.

With respect to electric power capacity, this entry proposes an alternative, and more practical, definition for sustainability, and also defines a purposefully unsustainable “commercial” capacity level (Fig. 2). The former is defined as the ability to economically maintain the installed capacity, over the amortized life of a power plant, by taking practical steps (such as, make-up well drilling) to compensate for resource degradation (pressure drawdown, well productivity decline and cooling). The latter can be defined as a capacity level that is initially kept higher than the sustainable level but may be allowed to decline with time once make-up well drilling, or other measures to mitigate resource degradation, becomes uneconomic at some point in project life. In a socially responsible vein, this declining capacity starting above the sustainable level could be considered commercial only if the levelized power cost is calculated to be lower than that from alternative renewable resources. Even if the power cost at such a commercial level proves higher than that from fossil fuels, this higher capacity can displace fossil fuel usage if power from renewable or environmentally benign resources is given adequate tax breaks (such as carbon credit), market access (such as implementation of



Geothermal Power Capacity, Sustainability and Renewability of. Figure 2
Proposed definitions [7]

“renewable energy portfolio standards”), or price support (such as production tax credit or any direct subsidy) by governments or international agencies.

The appropriate un-sustained but commercial power capacity level can only be arrived at by numerical simulation of the actual production behavior of the reservoir concerned and within the context of the economic realities and market forces. Such a purposefully un-sustained but commercial level is socially beneficial for a market-driven economy because it allows reduction in leveled power cost through accelerated capital recovery while helping to displace the use of fossil fuels. The cumulative energy extraction over the project life at an un-sustained but commercial level need not exceed the cumulative energy that would be extracted at the sustainable level, thus still assuring natural replenishment of the resource base in a 100–300-year time frame. Therefore, such a commercial development level is not only reasonable but also desirable, particularly if one considers the distinct possibility of acceleration of natural recharge of hot water into the reservoir, thus mitigating the impact of a higher initial production rate.

In discussing renewability and sustainability of geothermal energy, interesting analogies have been invoked from time to time by various authors, for example, comparison with mining, management of fisheries, utilization of hydropower, and so on. While all these analogies correspond to some aspects of geothermal energy exploitation, yet another analogy is offered here to elucidate the over-arching concept of sustainability proposed in this entry. A reasonable analogy for renewable capacity would be seasonal harvest of crops while timber harvest would be an appropriate analogy for sustainable capacity, for the timber resource would grow back within a few decades. One could harvest only the annual growth at the tips of the tree branches and keep the forest resource constantly renewable. But is this a reasonable approach to natural resource husbandry? While renewable, annual tree growth can be used as firewood or turned into paper pulp, the forest resource is more valuable to the society if mature trees are harvested for timber and then allowed to grow back. Likewise, constraining geothermal energy exploitation within a continuously renewable level, which is suitable primarily for low-intensity, non-electrical uses, is neither reasonable nor desirable from a socioeconomic viewpoint. In addition, thinning

of a forest accelerates tree growth due to the penetration of more sunlight into the forest; this is a convenient metaphor for the increase in natural recharge rate due to exploitation of a geothermal resource above the so-called renewable level.

Relation Between Renewable and Sustainable Capacities

This entry considers only liquid-dominated geothermal fields with capacity for supplying electric power, steam-dominated fields being rare occurrences; only six steam-dominated fields have been exploited to date: The Geysers, California; Lardarello, Italy; Matsukawa, Japan; Kamojang and Darajat; Indonesia; and Los Azufres, Mexico. Based on the experience in monitoring many producing geothermal fields for more than 3 decades and conducting dozens of numerical simulation studies of actual reservoirs, the author has observed that the sustainable capacity of a liquid-dominated field is typically an order of magnitude higher than the renewable capacity. The understanding here is that the renewable capacity of a field corresponds to the power capacity equivalent of the natural heat recharge, conductive plus convective, into the system; and sustainable capacity is supported by “mining” (or “harvesting” if one considers a time frame of centuries) of the stored heat in addition to natural heat recharge. To confirm this empirical observation, a review has been made of both published and unpublished results of numerical simulation and heat flow studies of more than half of the approximately 65 liquid-dominated geothermal fields in the world that have supplied commercial power to date and for which reasonably reliable estimate of the natural heat recharge rate could be made. The heat recharge rate was estimated from either numerical simulation of the reservoir or surface heat flow studies, with the reasonable assumption that the rate of natural heat recharge into the reservoir to be equal to the total rate of heat discharge at the surface over the entire thermal anomaly.

Table 1 lists approximate estimates of the renewable and sustainable capacities of 37 geothermal fields from published sources or various archives. The electrical power equivalent (MWe) was approximated from the estimated thermal power capacity based on First and Second Laws of Thermodynamics assuming a rejection

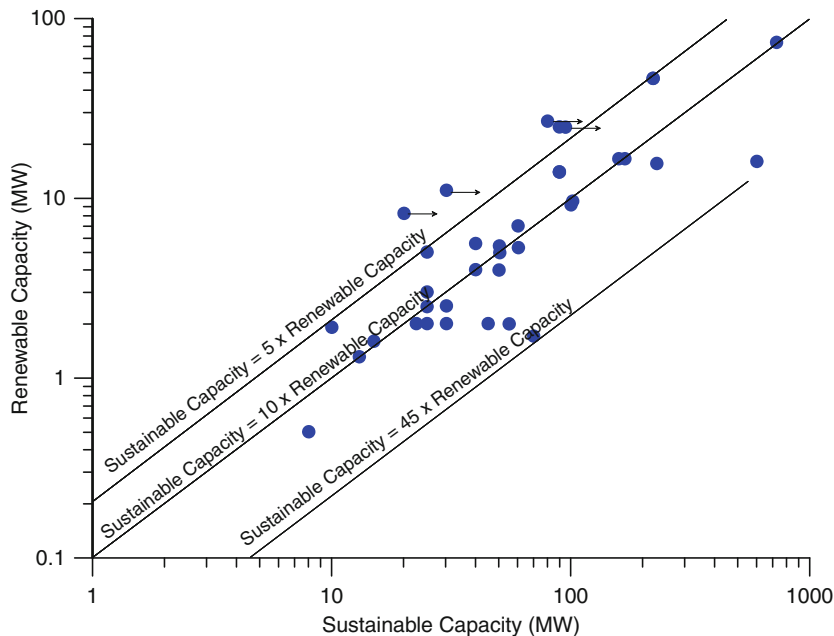
temperature of 15°C and a utilization factor of 0.45. The sustainable capacity value for a field in Table 1 was taken as the proven exploitation capacity, unless actual reservoir response and/or simulation studies had indicated the sustainable capacity to be higher. As such, the sustainable capacity values in Table 1 should in general be considered minimum estimates. As mentioned before, this table illustrates that renewable capacities are relatively small compared to sustainable capacities, the total for 37 fields being 386 and 2,056+ MWe, respectively. Furthermore, at the renewable level, most fields would not support commercial power development; for example, if 10 MWe were the smallest commercially developable capacity, only 11 of the 37 fields would qualify.

Figure 3 is a cross-plot of the above-listed renewable and sustainable capacities. The points with arrows in the direction of higher sustainable capacity represent fields for which the presently installed capacity appears manifestly smaller than the sustainable capacity but no estimate of the latter is available. This figure confirms the empirical observation that sustainable capacity is typically an order of magnitude higher than renewable capacity. Specifically, sustainable capacity (E_s) is

a multiple, Q , of renewable capacity (E_r), where α ranges from about 5–45, with a value of 10 most likely. The author has always observed that α , which can be termed the “Sustainability Factor,” tends to be high for a hydrothermal reservoir if the host formation is sedimentary. This is to be expected because having intergranular porosity, a porous sedimentary formation would display better heat transfer characteristics than a fractured non-sedimentary formation.

Wisian et al. [12] concluded from surface heat flow studies of a large number of geothermal fields that the presently installed capacity in most fields is equivalent to no more than ten times the natural heat discharge rate at the surface. Their conclusion at first seems to contradict this entry’s assumption that the sustainable capacity is 5–45 times the natural heat discharge rate, 10 times being most likely rather than the maximum. This difference can be explained by the fact that Wisian et al. [12] considered installed plant capacity, which is in general smaller than the maximum sustainable capacity.

Finally, the empirical observation that the sustainable capacity of a reservoir is an order of magnitude higher than the renewable capacity implies that,



Geothermal Power Capacity, Sustainability and Renewability of. Figure 3
Renewable capacity versus sustainable capacity [7]

following exploitation, the reservoir is expected to take an order of magnitude higher time span compared to the exploitation period for complete natural replenishment. This supports the earlier observation from reservoir simulation that the depletion effects of power production for 20–30 years would require on the order of 100–300 years to completely disappear.

Estimation of Renewable, Sustainable, and Commercial Capacities

The best tool for quantifying renewable capacity is a numerical simulation model that reproduces the natural physical state of the reservoir. But estimating sustainable and commercial capacities requires not only natural-state modeling but also trial-and-error matching of the actual exploitation history of the reservoir, and forecasting its behavior, using a reservoir simulation model. Estimation of an un-sustained commercial capacity also requires market considerations and economic analysis. Assessment of even renewable capacity may require trial-and-error history matching and forecasting if the recharge rate increases with reservoir pressure decline, which is sometimes the case. Obviously, the effective use of such numerical simulation requires adequate data on the natural state of the reservoir and significant production history. For some fields, renewable and sustainable capacities can be approximated by simple, “lumped-parameter” modeling of the production history. For many fields, data may not be available for numerical simulation or even for relatively simple lumped-parameter modeling. For such situations, approximate formulations to quantify these capacities are presented in Sanyal et al. [34] and are reproduced below.

By definition, Renewable Capacity (E_r) is given by [34]:

$$E_r = R = D_{\text{cond}}, \quad (1)$$

where R is heat recharge rate into the reservoir (primarily convective with a small conductive component) and D_{cond} is total heat discharge from the surface over the thermal anomaly; if the entire heat anomaly on the surface is considered, the convective component of heat discharge is usually negligible. Ideally, D_{cond} should be estimated from a comprehensive “heat budget” survey of the anomaly including conductive heat loss at the surface, convective heat discharge (through hot springs,

fumaroles and geysers) at surface manifestations, and subsurface convective heat loss to regional aquifers.

Strictly speaking, the small rate of background (regional) heat flow should be subtracted from the estimates of renewable capacity above and sustainable capacity as presented below [34]. However, given the approximate nature of such estimation, this correction is unnecessary in most situations.

Sustainable capacity (E_s), considering both heat mining and heat recharge, is given as [34]:

$$E_s = \left\{ \left(\frac{C_v}{KL} \right) rhd \left(\frac{A_{\text{res}}}{A} \right) + 1 \right\} D_{\text{cond}}, \quad (2)$$

where C_v is volumetric specific heat of fluid-filled rock, K is thermal conductivity of the overburden, L is plant life, r is heat energy recovery factor, h is reservoir thickness, d is depth to the top of the reservoir, A_{res} is reservoir area, and A is the area of the entire thermal anomaly.

A conservative definition of commercial capacity (E_c) would require that $E_c > E_s$ initially, but eventually falls below E_s , such that the total energy recovered over the plant life is same as that would be for production at the sustainable level. With this definition, and “harmonic decline” in well productivity, it can be shown [34]:

$$E_c = \frac{E_s L D_i}{\ln(1 + D_i L)}, \quad (3)$$

where D_i is initial decline rate in well productivity. E_c can be considerably higher than E_s , depending on economic factors. The higher the margin by which E_c exceeds E_s , the higher is D_i .

An actual example, that of the Beowawe geothermal field in the State of Nevada, United States, can be considered. For this field,

$$\left(\frac{A_{\text{res}}}{A} \right) \approx 0.1$$

$$d = 900 \text{ m, and } h = 1,500 \text{ m}$$

From Butler et al. [10], for this field, $R = 1.3 \text{ MWe} \approx D_{\text{cond}}$ (ignoring background heat flow).

Therefore, Renewable Capacity = 1.3 MWe

Typical values of the other parameters are: $C_v = 2,700 \text{ kJ/m}^3/\text{°C}$, $K = 3.1 \text{ W/m/°C}$, $L = 30 \text{ years}$ and $r = 0.1$.

Therefore, from Eq. 2, Sustainable Capacity $\approx 18 \text{ MWe}$ (ignoring background heat flow).

Most likely reserves for this field, from Klein et al. [35] = 58 MWe.

Therefore, commercial capacity would fall somewhere between 18 and 58 MWe, depending on the economic factors. For example, if no make-up well drilling is contemplated and an initial harmonic productivity decline rate of 10% is economically acceptable, from Eq. 3, $E_c = 40$ MWe.

The above discussion shows that the renewable development level for the Beowawe field is only 1.3 MWe, which is entirely uneconomic. While a sustainable capacity of 18 MWe is commercial, a capacity of 40 MWe may even be more attractive economically, and yet would cause no further cumulative energy withdrawal from the reservoir over a 20–30-year project life, and consequently, the reservoir should still be replenished naturally, in a 100–300-year time frame. It should be noted that a plant capacity of 13 MWe has already been sustained in this field over the past 2 decades.

An Illustrative Case History

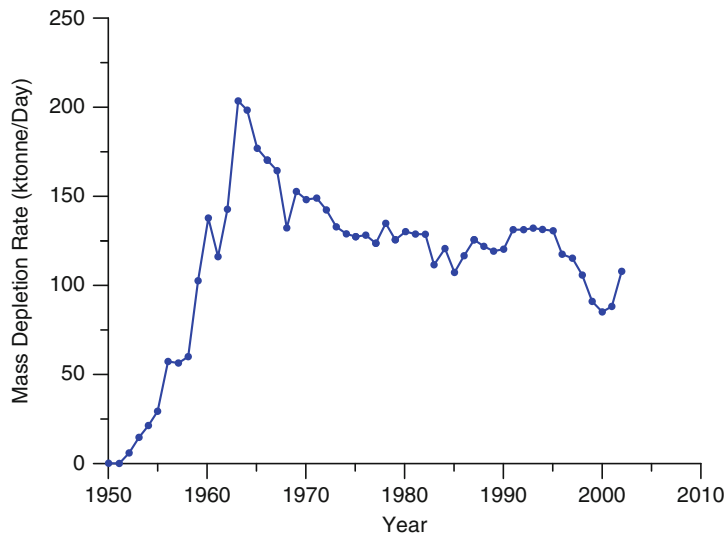
This is a case history of estimating renewable and sustainable capacities of a geothermal field (at Wairakei, New Zealand) from its production history using a simple “lumped-parameter” model. The Wairakei field presents a good case history because: (a) it has more than 50 years of production history, longer than

that of any other liquid-dominated field in the world; (b) it offers an extensive database that is publicly available (for example, Clotworthy [36]); and (c) since the average temperature of this reservoir has not declined significantly over its long production history, its pressure behavior can be reasonably modeled by considering material balance only (rather than coupled material-and-energy balance).

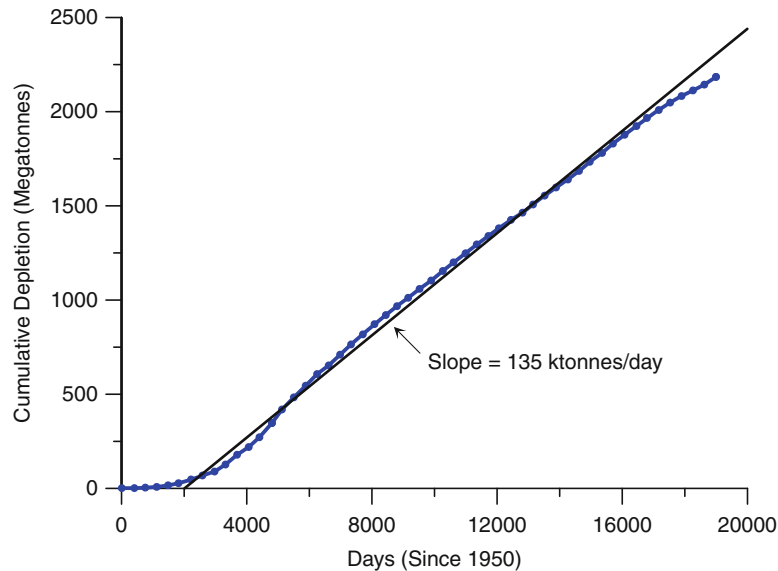
Numerical simulation and heat flow studies of this field have shown the steady-state recharge rate in the natural state to be about 31 kt/day; in other words, the minimum renewable depletion capacity (E_r) is 31 kt/day. Figure 4 presents a plot of the mass depletion rate (m), defined as production rate minus injection rate, versus time at this field. As of 1956 (2,000 days from the initiation of production in 1950), the reservoir pressure in the deep liquid zone in the Western Borefield (the portion of the field eventually most exploited) was about 52 bar-a, and negligible production had taken place before that time. From material balance consideration, it can be shown [7] that reservoir pressure (p) is given as:

$$p = 52.0 - \frac{(m - 31)}{r} \left[1 - e^{-\frac{r}{s}(t-2000)} \right], \quad (4)$$

where m is assumed constant with time (t , days), r is a recharge coefficient (kt/day/bar), and s is a reservoir storage coefficient (kt/bar).



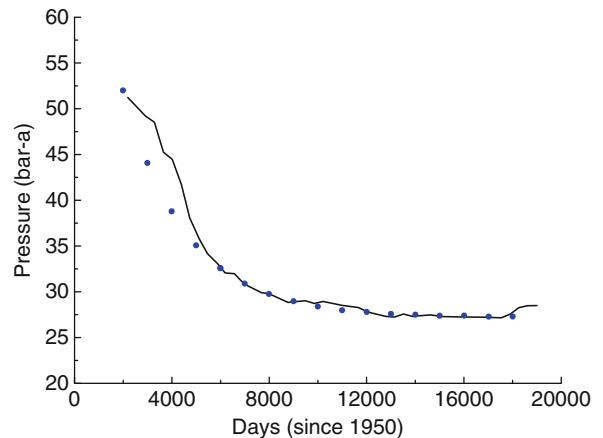
Geothermal Power Capacity, Sustainability and Renewability of. Figure 4
Mass depletion history [7]



Geothermal Power Capacity, Sustainability and Renewability of. Figure 5
Cumulative depletion history [7]

Figure 5 shows the cumulative depletion history of the field. Between 2,000 days and the present, a reasonably linear trend can be defined with a slope of 135 kt/day. Therefore, one can approximate a constant value of m after 2,000 days as 135 kt/day. The unknowns r and s in Eq. 4 can be estimated by trial-and-error; Fig. 6 shows the best fit the author obtained between the observed pressure (continuous curve) at the deep liquid zone of the Western Borefield and the computed pressures (solid circles) as a function of time; this fit required an s value of 11,000 kt/bar and an r value of 4.2 kt/day/bar. The fit in Fig. 6 is good between 5,000 and 18,000 days, a span of 36 years; a look at Fig. 5 shows that the poor match before and after this period is to be expected as the depletion trend had deviated significantly from the linear in the very early and very recent periods.

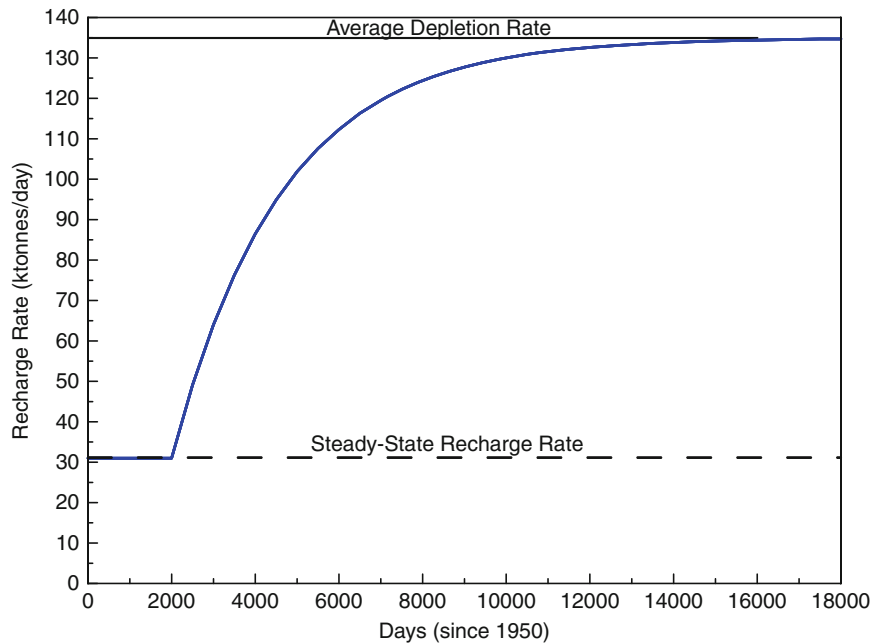
The overall recharge rate at any time is the sum of the steady-state recharge rate (m_{ss}) and the pressure-dependent component of recharge rate (m_r) [7]. Using the r and s values derived above, the historical rate of recharge at Wairakei has been estimated as shown in Fig. 7. Overall, fluid recharge at Wairakei to date appears to have been generally hot because negligible overall cooling of the reservoir has been noted in 50



Geothermal Power Capacity, Sustainability and Renewability of. Figure 6

Observed and computed liquid pressures, western borefield [7]

years, and recharge has steadily increased in response to pressure drawdown (Fig. 7). For this reason, the renewable level of depletion of this reservoir has become steadily higher than the steady-state depletion rate of 31 kt/day derived from natural-state modeling. In fact,



Geothermal Power Capacity, Sustainability and Renewability of. Figure 7
Recharge rate versus time, Wairakei field [7]

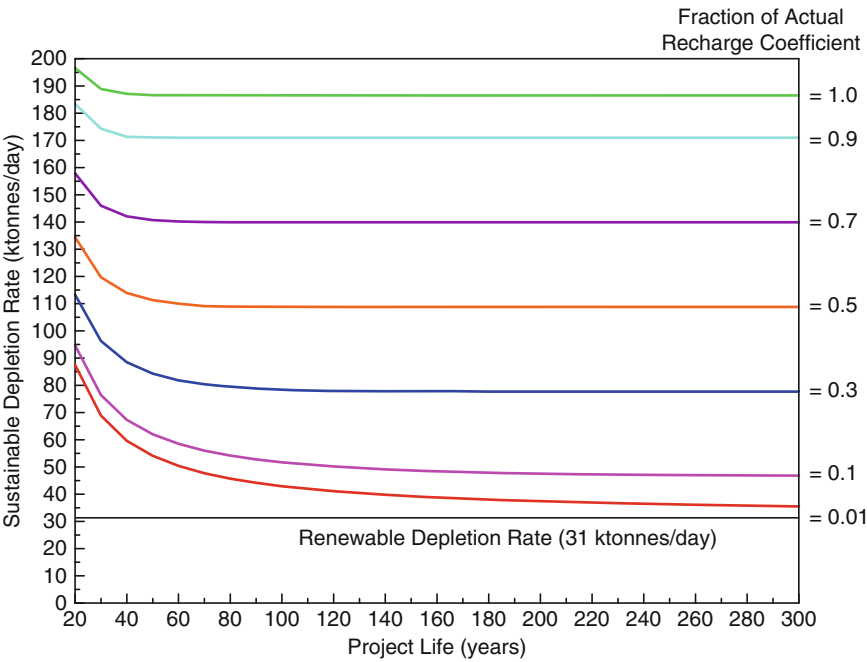
the recharge rate by 17,000 days has nearly equaled the depletion rate; if the entire recharge here indeed represents hot fluid entry from depth, then a depletion level of 135 kt/day, rather than 31, can be considered renewable.

Now, what is the sustainable depletion capacity (E_s) of this reservoir? If the minimum static reservoir pressure at which wells in this field can still flow commercially can be estimated, then one can calculate E_s for any assumed project life. Wellbore simulation for wells producing from the deep liquid zone at Wairakei indicates this minimum pressure value to be about 15 bar-a. Therefore, one can calculate the sustainable capacity, assuming only hot recharge, for any assumed project life from Eq. 4.

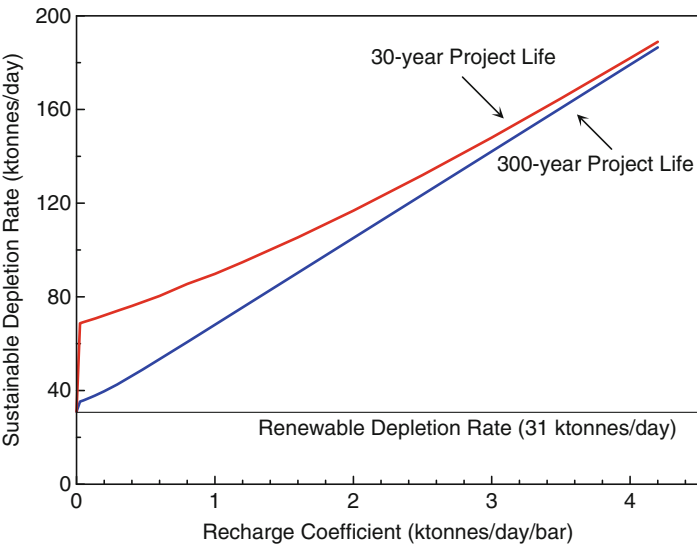
The above equation gives very similar values of E_s for a 30-year project life or a 300-year project life, 188.8 and 186.5 kt/day, respectively. This relative insensitivity of E_s to project life is due to the very high recharge coefficient and the apparent preponderance to date of hot rather than cool recharge at Wairakei; this latter fact is also supported by numerical simulation. Since recharge rate in most fields is lower

than at Wairakei, an assessment was made of how E_s would have changed as a function of project life if the recharge coefficient at Wairakei were smaller. Figure 8 shows the calculated E_s value versus project life for a range of hypothetical recharge coefficients expressed as fractions of the actual recharge coefficient at Wairakei. Figure 8 shows that as the recharge coefficient becomes smaller, so does sustainability and the latter becomes more sensitive to project life. Figure 9 shows the same data as in Fig. 8 represented as sustainable capacity versus recharge coefficient for project lives of both 30 and 300 years. This figure illustrates that the difference between renewable and sustainable capacities for 30- and 300-year project lives becomes less as recharge coefficient increases, for Wairakei this difference (corresponding to an r of 4.2 kt/day) being negligible.

Finally, it should be noted that sustainability factor (φ), as defined before, for Wairakei is 188.8/31, or 6.1. Why is this value of sustainability factor at the low end of the range of 5–45 mentioned earlier? The reason is that until recently, there was no injection in this field. Therefore, the above analysis is based on



Geothermal Power Capacity, Sustainability and Renewability of. Figure 8
Sustainable depletion rate versus project life, Wairakei field [7]



Geothermal Power Capacity, Sustainability and Renewability of. Figure 9
Sustainable capacity versus recharge coefficient [7]

depletion being equal to production. If injection is practiced, the effective depletion rate will be lower than production rate, and therefore, a higher production capacity can be sustained. For example, if

50% of the produced fluid were injected, the sustainable production rate would be double the sustainable depletion rate (188.8 kt/day), that is, 377.6 kt/day, assuming the recharge to be predominately hot.

Therefore, the sustainability factor would be $377.6/31$ or 12.2 ; this sustainable production capacity is an order of magnitude higher than the renewable capacity of 31 kt/day.

Future Directions

The debate over renewability and sustainability of power capacity of a geothermal reservoir still continues for conventional geothermal systems, which are naturally occurring subsurface porous or fractured systems that can be tapped for production by drilling wells. However, in the last decade, considerable hopes have been raised of tapping geothermal energy from enhanced geothermal systems (“EGS”) [37]. These are hot subsurface systems with porosity or fracture capacity too low to allow commercial production but can be enhanced by pervasive hydraulic fracturing to enable significant fluid injection and production. In an EGS project, heat is recovered from the artificial reservoir by injecting cool water through a set of wells while producing heated water from another set of wells. Such systems have not yet proven commercial, but research and development toward commercial tapping of EGS systems continue.

If EGS systems can be exploited commercially, the energy reserves in such systems in the USA would be two orders of magnitude larger than the energy contained in the conventional geothermal systems [38]. Even in countries where conventional geothermal systems do not exist, EGS developments would be the theoretically possible, because anywhere on earth adequately hot rock bodies can be reached by drilling wells deep enough and creating an artificial reservoir by hydraulic fracturing of rock.

Renewability and sustainability of EGS systems have not received much attention yet, but in the future, this issue will become important if EGS exploitation becomes a commercial reality. The one major difference between renewability and sustainability of conventional systems and those of an EGS is that an EGS reservoir does not receive natural convective heat recharge; all heat recharge to an EGS reservoir would be conductive which, as discussed earlier, is relatively minor. Furthermore, an abandoned EGS project would

not be fully replenished in 100–300 years as expected for conventional systems because of this lack of convective heat recharge.

Bibliography

1. Axelsson G, Stefansson V, Björnsson G (2004) Sustainable utilization of geothermal resources. In: Proceedings of the twenty-ninth workshop on geothermal reservoir engineering, Stanford University, Stanford, pp 26–28
2. Rybach L (2003) Sustainable use of geothermal resources: renewability aspects. In: IGC 2003 short course, UNU Geothermal Training Programme, Iceland
3. Axelsson G, Gudmundsson A, Steingrímsson B, Palmason G, Armannsson H, Tulinius H, Flovenz OG, Björnsson S, Stefansson V (2001) Sustainable production of geothermal energy: suggested definition. IGA-News Quarterly No. 43:1–2
4. Stefansson V (2000) The renewability of geothermal energy. In: Proceedings World geothermal congress, Kyushu-Tohoku, Japan, 28 May–10 June 2000
5. Rybach L, Mégel T, Eugster WJ (1999) How renewable are geothermal resources? Trans Geotherm Res Council 23:563–566
6. Wright PM (1995) The sustainability of production from geothermal resources. In: Proceedings of the World geothermal congress, Florence, Italy, 18–31 May 1995
7. Sanyal S (2005) Sustainability and renewability of geothermal power capacity. In: Proceedings of the World geothermal congress, Antalya, Turkey, 24–29 Apr 2005
8. Pritchett JW (1998) Modeling post-abandonment electrical capacity recovery for a two-phase geothermal reservoir. Trans Geotherm Res Council 22:521–528
9. Parini M, Cappetti G, Laudiano M, Bertani R, Monterrosa M (1995) Reservoir modeling study modeling study of the Ahuachapán geothermal field (El Salvador) in the frame of a generation stabilization project. In: Proceedings of World geothermal congress, Florence, Italy, 18–31 May 1995
10. Butler SJ, Sanyal SK, Robertson-Tait A, Lovekin JW, Benoit D (2001) A case history of numerical modeling of a fault-controlled geothermal system at Beowawe, Nevada. In: Proceedings of the twenty-sixth workshop on geothermal reservoir engineering, Stanford University, Stanford, 29–31 Jan 2001
11. Butler SJ, Sanyal SK, Henneberger RC, Klein CW, Gutiérrez H, de León JS (2000) Numerical modeling of the Cerro Prieto geothermal field, Mexico. In: Proceedings of the World Geothermal Congress, Kyushu-Tohoku, Japan, 28 May–10 June 2000
12. Wisian KW, Blackwell DD, Richards M (2001) Correlation of surface heat loss and total energy production for geothermal systems. Trans Geotherm Res Council 25:331–336
13. Lippmann MJ, Bodvarsson GS (1985) The Heber geothermal field, California: natural state and exploitation modeling studies. J Geophys Res 90(B1):745–758
14. McGuinness M, White S, Young R, Ishizaki H, Ikeuchi K, Yoshida Y (1995) A model of the Kakkonda geothermal reservoir. Geothermics 24:1–48

15. White SP, Kissling WM, McGuinness MJ (1997) Models of the Kawareu geothermal reservoir. *Trans Geotherm Res Council* 21:33–39
16. Tulinius H, Sigurdsson O (1989) Two-dimensional simulation of the Krafla-Hvitholar geothermal field, Iceland. In: *Proceedings of the fourteenth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 24–26 Jan 1989
17. Sorey ML (1985) Evolution and present state of the hydrothermal system in the Long Valley caldera. *J Geophys Res* 90:11219–11228
18. Esberto MB, Sarmiento ZF (1999) Numerical modeling of the Mt. Apo geothermal reservoir. In: *Proceedings of the twenty-fourth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 25–27 Jan 1999
19. Haukwa C, Bodvarsson GS, Lippmann MJ, Mainieri A (1992) Preliminary reservoir engineering studies of the Miravalles geothermal field, Costa Rica. In: *Proceedings of the seventeenth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 29–31 Jan 1991
20. Sakagawa Y, Takahashi M, Hanano M, Ishido T, Demboya N (1994) Numerical simulation of the Mori geothermal field, Japan. In: *Nineteenth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 18–20 Jan 1994
21. Kiryukhin AV (2004) Modeling study of the Mutnovsky geothermal field (Dachny) in connection with the problem of steam supply for 50 MWe PP. In: *Twenty-ninth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 26–28 Jan 2004
22. Steingrímsson B, Bodvarsson GS, Gunnlaugsson E, Gislason G, Sigurdsson O (2000) Modeling studies of the Nesjavellir geothermal field, Iceland. In: *Proceedings of the World geothermal congress, Kyushu-Tohoku, Japan, 28 May–10 June 2000*
23. McGuinness MJ (1998) Ngawha geothermal field – a review. In: *Proceedings of the twentieth New Zealand geothermal workshop*, University of Auckland, Auckland, New Zealand
24. Yamada M, Iguchi K, Nakanishi S, Todaka N (2000) Reservoir characteristics and development plan of the Oguni geothermal field, Kyushu, Japan. *Geothermics* 29:151–169
25. Nakanishi S, Nobuyuki I (2000) Reservoir simulation study of the Onikobe geothermal field, Japan. In: *Proceedings of the World geothermal congress, Kyushu-Tohoku, Japan, 28 May–10 June 2000*
26. Yearsley E (1994) Roosevelt hot springs reservoir model applied to forecasting remaining field potential. *Trans Geotherm Res Council* 18:617–622
27. Atmojo JP, Itoi R, Fukuda M, Tanaka T, Daud Y, Sudarman S (2001) Numerical modeling study of Sibayak geothermal reservoir, North Sumatra, Indonesia. In: *Proceedings of the twenty-sixth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 29–31 Jan 2001
28. Pritchett JW, Garg SK, Ariki K, Kawano Y (1991) Numerical simulation of the Sumikawa geothermal field in the natural state. In: *Proceedings of the sixteenth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 23–25 Jan 1991
29. Furuya S, Aoki M, Gotoh H, Takenaka T (2000) Takigami geothermal system, Northeastern Kyushu, Japan. *Geothermics* 29:191–211
30. Butler SJ, Sanyal SK, Klein CW, Iwata S, Itoh M (2004) Numerical simulation and performance evaluation of the Uenotai geothermal field, Akita Prefecture. *Jpn Trans Geotherm Res Council* 28:455–460
31. Bibby HM, Caldwell TG, Davey FJ, Webb TH (1995) Geophysical evidence on the structure of the Taupo volcanic zone and its hydrothermal circulation. *J Volcanol Geotherm Res* 68:29–58
32. Sanyal SK, Pham M, Iwata S, Suzuki M (2000) Numerical simulation of the Wasabizawa geothermal field, Akita Prefecture. *Jpn Trans Geotherm Res Council* 24:623–630
33. Menzies AJ, Granados EE, Sanyal SK, Mérida-I L, Caicedo AA (1991) Numerical modeling of the initial state and matching of well test data from the Zunil geothermal field, Guatemala. In: *Proceedings of the sixteenth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 23–25 Jan 1991
34. Sanyal SK, Klein CW, Lovekin JW, Henneberger RC (2004) National assessment of U.S. geothermal resources – a perspective. *Trans Geotherm Res Council* 28:355–362
35. Klein CW, Lovekin JW, Sanyal SK (2004) New geothermal site identification and quantification. In: *California energy commission PIER consultant report, P500-04-051*
36. Clotworthy A (2000) Response of Wairakei geothermal reservoir to 40 years of production. In: *Proceedings of the World geothermal congress, Kyushu-Tohoku, Japan, 28 May–10 June 2000*
37. MIT (2006) The future of geothermal energy – impact of enhanced geothermal systems (EGS) on the United States in the 21st century. An assessment by an MIT – Led interdisciplinary panel, Massachusetts Institute of Technology, Cambridge
38. Sanyal SK (2010) Future of geothermal energy. In: *Proceedings of the thirty-fifth workshop on geothermal reservoir engineering*, Stanford University, Stanford, 1–3 Feb 2010, SGP-TR-188

Geothermal Power Conversion Technology

LUCIEN Y. BRONICKI

Ormat Technologies, Inc., Reno, NV, USA

Article Outline

Glossary

Definition of Geothermal Power Conversion Technology

Introduction

Geothermal Project Design and Implementation

Geothermal Resources

Thermodynamic Analysis of the Energy Conversion Process
 Main Power Station Components
 Choosing the Energy Conversion Systems
 Commercial Power Stations
 Organic Rankine Cycle Configurations for Geothermal Power Stations
 Experimental Power Stations
 Future of Geothermal Energy
 Acknowledgments
 Bibliography

Glossary

Ambient Natural condition of the environment at any given time.

Baseload The lowest level of power production needs during a season or year.

Baseload plants Electricity-generating units that are operated to meet the constant or minimum load on the system. The cost of energy from such units is usually the lowest available to the system.

Binary-cycle power station A geothermal electricity-generating station employing a closed-loop heat exchange system in which the heat of the geothermal fluid (the “primary fluid”) is transferred to a different fluid (“motive,” “secondary,” or “working” fluid), which is thereby vaporized and used to drive a turbine/generator set.

Bi-phase expander A bi-phase expander or bi-phase turbine is a device that produces power by utilizing the energy of two-phase (liquid/vapor) streams. The total energy produced by the brine and the separated steam (in a reduced condition), is expected to be higher than that of the steam turbine alone.

Brine A geothermal solution containing appreciable amounts of sodium chloride or other salts.

Capacity factor A percentage that tells how much of a power station’s capacity is used over time. For example, typical station capacity factors range as high as 80% for geothermal and 70% for cogeneration.

Capacity, installed (or Nameplate) The total manufacturer-rated capacities of equipment such as turbines, generators, condensers, transformers, and other system components.

Capacity The amount of electric power delivered or required for which a generator, turbine, transformer,

transmission circuit, station, or system is rated by the manufacturer.

Carbon dioxide A colorless, odorless, nonpoisonous gas that is a normal part of the air. Carbon dioxide, also called CO₂, is exhaled by humans and animals and is absorbed by green growing things and by the sea.

CHP Combined heat and power

Condensate Liquid formed by condensation of vapor.

Cooling tower A structure in which process heat is removed to the atmosphere.

Cost The amount paid to acquire resources, such as station and equipment, fuel, or labor services.

Demand The level at which electricity or natural gas is delivered to users at a given point in time. Electric demand is expressed in kilowatts.

Direct use Use of geothermal heat without first converting it to electricity, such as for space heating and cooling, food preparation, industrial processes, etc.

Dispatch The operating control of an integrated electric system to assign generation to specific generating stations and other sources of supply to effect the most reliable and economical supply as the total of the significant area loads rise or fall. Control operations and maintenance of high-voltage lines, substations and equipment, including administration of safety procedures. Operate the interconnection. Schedule energy transactions with other interconnected electric utilities.

Drift eliminator Drift eliminators reduce the amount of drift in the exiting air flow. Drift droplets can be reduced to less than 0.1% by effective use of an eliminator.

Drift Drift droplets are any water droplets and dissolved and suspended solids that are entrained in the air and emitted from the cooling tower stack.

Dry steam Very hot steam that does not occur with liquid.

Efficiency The ratio of the useful energy delivered by a dynamic system (such as a machine, engine, or motor) to the energy supplied to it over the same period or cycle of operation. The ratio is usually determined under specific test conditions.

Effluent Treated wastewater.

EGS Engineered geothermal systems

Emissions standard The maximum amount of a pollutant legally permitted to be discharged from a single source.

Energy efficiency Refers to programs that are aimed at reducing the energy used by specific end-use devices and systems, typically without affecting the services provided. These programs reduce overall electricity consumption (reported in megawatt-hours), often without explicit consideration for the timing of program-induced savings. Such savings are generally achieved by substituting technically more advanced equipment to produce the same level of end-use services (lighting, heating, and motor drive) with less electricity. Examples include high-efficiency appliances, efficient lighting programs, high-efficiency heating, ventilating and air conditioning (HVAC) systems or control modifications, efficient building design, advanced electric motor drives, and heat recovery systems.

Energy source The primary source that provides the power that is converted to electricity through chemical, mechanical, or other means. Energy sources include coal, petroleum and petroleum products, gas, water, uranium, wind, sunlight, and geothermal and other sources.

Energy The capacity for doing work as measured by the capability of doing work (potential energy) or the conversion of this capability to motion (kinetic energy). Energy has several forms, some of which are easily convertible and can be changed to another form useful for work. Most of the world's convertible energy comes from fossil fuels that are burned to produce heat that is then used as a transfer medium to mechanical or other means in order to accomplish tasks. Electrical energy is usually measured in kilowatt-hours, while heat energy is usually measured in British thermal units.

Facility An existing or planned location or site at which prime movers, electric generators, and/or equipment for converting heat into electric energy are situated, or will be situated. A facility may contain more than one generator of either the same or different prime mover type.

Fault A fracture or fracture zone in the Earth's crust along which slippage of adjacent Earth material has occurred at some time.

Flash steam Steam produced when the pressure on a geothermal liquid is reduced. Also called flashing.

Flash tank Vessel in which the geothermal water or brine is flashed into steam by pressure reduction.

Fly ash Particulate matter from coal ash in which the particle diameter is less than 1×10^{-4} m. This is removed from the flue gas using flue gas particulate collectors such as fabric filters and electrostatic precipitators.

Generation (electricity) The process of producing electric energy by transforming other forms of energy also, the amount of electric energy produced, expressed in watt-hours (Wh).

Generator A machine that converts mechanical energy into electrical energy.

Geology Study of the planet Earth, its composition, structure, natural processes, and history.

Geothermal combined cycle An electricity-generating technology in which electricity is produced from the steam exiting from one or more steam turbines at above atmospheric pressure. The exiting steam routed to the evaporator of an ORC station producing electricity. This process reduces the impact of non-condensable gases in the geothermal steam and eliminates the power consumption of the vacuum pumps or ejectors (updated using various sources).

Geothermal energy Natural heat from within the Earth, captured for production of electric power, space (geofluid) heating or industrial steam.

Geothermal fluid Can be steam, water, brine or a mixture of two, may contain noncondensable gases (CO_2 , H_2S) and in case of brine, appreciable amounts of sodium chloride, carbonates, and silica.

Geothermal heat pumps Devices that take advantage of the relatively constant temperature of the Earth's interior, using it as a source and sink of heat for both heating and cooling. When cooling, heat is extracted from the space and dissipated into the Earth when heating, heat is extracted from the Earth and pumped into the space.

Geothermal power station A power station in which the prime movers are turbines operated either by steam or organic fluids vapor. The steam is either natural or produced from flashing of hot water. The organic fluid vapor is produced by boiling of the organic fluid using geothermal steam or water.

The natural steam and water derive energy from heat found in rocks or fluids at various depths beneath the surface of the Earth. The energy is extracted by drilling and/or pumping. It includes all the surface facilities including the geothermal fluid gathering and treatment system, but does not include the geothermal wells and pumps.

Geothermal steam Steam drawn from deep within the Earth.

Geothermal Of or relating to the Earth's interior heat.

Geyser A spring mat that shoots jets of hot water and steam into the air.

Geysers, The A large geothermal steam field located approximately 75 miles (121 km) north of the city of San Francisco, California.

Gigawatt (GW) One billion watts.

Gigawatt-hour (GWh) One billion watt-hours.

Greenhouse effect The increasing mean global surface temperature of the Earth caused by gases in the atmosphere (including carbon dioxide, methane, nitrous oxide, ozone, and chlorofluorocarbon). The greenhouse effect allows solar radiation to penetrate but absorbs the infrared radiation returning to space.

Grid The layout of an electrical distribution system.

Heat exchanger A device for transferring thermal energy from one fluid to another.

Hot dry rock (HDR) A geothermal resource created with impermeable, subsurface hot rock structures, typically granite rock below the Earth's surface. The resource is being investigated as a source of energy production.

Hybrid geothermal cycles Cycles in which there are in series or in parallel a steam Rankine cycle and an Organic Rankine cycle.

Hybrid geothermal power stations Stations in which the geothermal heat is supplemented by another heat source.

Hydrothermal resource Underground systems of hot water and/or steam.

Injection The process of returning spent geothermal fluids to the subsurface. Sometimes referred to as reinjection.

Kilowatt (kW) One thousand watts.

Kilowatt-hour (kWh) One thousand watt-hours.

Known geothermal resource area A region identified by the US Geological Survey as containing geothermal resources.

Leaching The removal of readily soluble components, such as chlorides, sulfates, organic matter, and carbonates, from soil by percolating water. The remaining upper layer of leached soil becomes increasingly acidic and deficient in plant nutrients.

Load (electric) The amount of electric power delivered or required at any specific point or points on a system. The requirement originates at the energy-consuming equipment of the consumers.

Magma The molten rock and elements that lie below the Earth's crust. The heat energy can approach 550°C and is generated directly from a shallow molten magma resource and stored in adjacent rock structures. To extract energy from magma resources requires drilling near or directly into a magma chamber and circulating water down the well in a convection-type system.

Megawatt (MW) One thousand kilowatts (1,000 kW) or one million watts (1,000,000 W).

Megawatt-hour (MWh) One million watt-hours.

Muffler It is a device for reducing noise of high-speed steam flow in emergency relief of high-pressure steam from the production well (PW) in the power station. In geothermal applications it not only acts as a silencer but also performs safety and environmental duties. This is because of the high temperature and high salinity of the steam and brine that is released to the atmosphere in case of turbine trip-off or system emergency shutdown.

Noncondensable gases (NCG) Gases present in the steam or dissolved in the brine and liberated in the flash process.

Ormat energy converter (OEC) A unit using Ormat's Organic Rankine Cycle technology, which converts geothermal heat to electric power.

ORC power station A power station operating according to the ORC process.

Organic Rankine cycle (ORC) A Rankine cycle using an organic fluid (updated using various sources).

Outage The period during which a generating unit, transmission line, or other facility is out of service.

Particulate matter (PM) Unburned fuel particles that form smoke or soot and stick to lung tissue when inhaled. A chief component of exhaust emissions from heavy-duty diesel engines

pH The term pH is a measure of acidity or alkalinity and ranges from 0 to 14. A pH measurement of 7 is

regarded as neutral. Measurements below 7 indicate increased acidity, while those above indicate increased alkalinity.

Point source A stationary location or fixed facility from which pollutants are discharged.

Power plant A power station.

Power station A facility at which prime movers electric generators, and auxiliary equipment are located, for converting mechanical, chemical, and/or nuclear energy into electric energy. A station may contain more than one type of prime mover.

Power Electricity for use as energy.

Precipitation Precipitation is the formation of a solid in a solution. The solid formed is called the precipitate, and the liquid remaining above the solid is called the supernate.

Price The amount of money or consideration-in-kind for which a service is bought, sold, or offered for sale.

Purifier Vessel at the turbine in which fine droplets are separated from the vapor.

Regulation The governmental function of controlling or directing economic entities through the process of rulemaking and adjudication.

Reliability Electric system reliability has two components adequacy and security. Adequacy is the ability of the electric system to supply to aggregate electrical demand and energy requirements of the customers at all times, taking into account scheduled and unscheduled outages of system facilities. Security is the ability of the electric system to withstand sudden disturbances, such as electric short circuits or unanticipated loss of system facilities. The degree of reliability may be measured by the frequency, duration, and magnitude of adverse effects on consumer services.

Renewable energy Resources that constantly renew themselves or that are regarded as practically inexhaustible. These include solar, wind, geothermal, hydro, and wood. Although particular geothermal formations can be depleted, the natural heat in the Earth is a virtually inexhaustible reserve of potential energy. Renewable resources also include some experimental or less-developed sources such as tidal power, sea currents, and ocean thermal gradients.

Renewable resources Natural but flow-limited resources that can be replenished. They are virtually inexhaustible in duration but limited in the amount

of energy that is available per unit of time. Some (such as geothermal and biomass) may be stock-limited in that stocks are depleted by use, but on a time scale of decades, or perhaps centuries, they can probably be replenished. Renewable energy resources include: biomass, hydro, geothermal, solar, and wind. In the future, they could also include the use of ocean thermal, wave, and tidal action technologies. Utility renewable resource applications include bulk electricity generation, on-site electricity generation, distributed electricity generation, non-grid-connected generation, and demand-reduction (energy efficiency) technologies.

Reservoir A natural underground container of liquids, such as water or steam (or, in the petroleum context, oil or gas).

Revenue The total amount of money received by a firm from sales of its products and/or services, gains from the sales or exchange of assets, interest and dividends earned on investments, and other increases in the owner's equity except those arising from capital adjustments.

Saturation Saturation is the point at which a solution of a substance can dissolve no more of that substance and additional amounts of it will appear as a precipitate. This point of maximum concentration, the saturation point, depends on the temperature of the liquid as well as the chemical nature of the substances involved.

Scaling Scaling is formation of a deposit layer (scale) on a solid surface, i.e., evaporators, pipes, etc.

Screw expander The screw expander is the reverse usage of a screw compressor consisting of two helical rotating wheels compressing gas in between them. When high-pressure gas is introduced to the compressor exit, it expands forcing the screw wheels to rotate backward and produce power.

Scrubber Equipment used to remove sulfur oxides or hydrogen sulfide from the geothermal fluid before discharge to the atmosphere. Chemicals, such as lime, are used as the scrubbing media. The scrubber is also used when fresh water is applied to saline-contaminated steam. The scrubber reduces the steam salinity before it enters the turbine.

Separator A vessel at the wellhead where steam is separated from water or brine. Mostly of centrifugal type.

Solubility Solubility is the property of a solid, liquid, or gaseous chemical substance called solute to dissolve in a liquid solvent to form a homogeneous solution of the solute in the solvent. The solubility of a substance fundamentally depends on the used solvent as well as on temperature and pressure.

Stability The property of a system or element by virtue of which its output will ultimately attain a steady state. The amount of power that can be transferred from one machine to another following a disturbance. The stability of a power system is its ability to develop restoring forces equal to or greater than the disturbing forces so as to maintain a state of equilibrium.

Steam Rankine cycle A Rankine cycle in which water (in liquid and vapor phase) is the motive fluid (updated using various sources).

System (electric) Physically connected generation, transmission, and distribution facilities operated as an integrated unit under one central management, or operating supervision.

System A combination of equipment and/or controls, accessories, interconnecting means, and terminal elements by which energy is transformed to perform a specific function, such as climate control, service water heating, or lighting.

Thermal pollution A reduction in water quality caused by increasing its temperature, often due to disposal of waste heat from industrial, power generation processes, or urban impervious surfaces (such as parking lots). Thermally polluted water can harm the environment because plants and animals may have difficulty adapting to it.

Transmission The movement or transfer of electric energy over an interconnected group of lines and associated equipment between points of supply and points at which it is transformed for delivery to consumers, or is delivered to other electric systems. Transmission is considered to end when the energy is transformed for distribution to the consumer.

Turbine generator A device that uses steam, organic vapor, heated gases, water flow or wind to cause spinning motion that activates electromagnetic forces and generates electricity (updated using various sources).

Turbine A machine for generating rotary mechanical power from the energy of expansion of a stream of

fluid (such as water, steam, organic vapor, or hot gas). Turbines convert the kinetic energy of fluids to mechanical energy through the principles of impulse and reaction, or a mixture of the two (updated using various sources).

Utility A regulated entity which exhibits the characteristics of a natural monopoly. For the purposes of electric industry restructuring, “utility” refers to the regulated, vertically integrated electric company. “Transmission utility” refers to the regulated owner/operator of the transmission system only. “Distribution utility” refers to the regulated owner/operator of the distribution system which serves retail customers.

Vapor(or steam)-dominated resources A geothermal reservoir system in which subsurface pressures are controlled by vapor rather than liquid and most of the flow is steam.

Water-dominated resource A resource where the major part of the mass flow is water or brine.

Watt The electrical unit of power. The rate of energy transfer equivalent to 1 A flowing under a pressure of 1 V at unity power factor.

Watt-hour (Wh) An electrical energy unit of measure equal to 1 W of power supplied to, or taken from, an electric circuit steadily for 1 h.

Definition of Geothermal Power Conversion Technology

Geothermal Power Conversion Technology refers to techniques used for the conversion of the heat content of geothermal fluid into mechanical power in order to drive a generator and produce electric power.

The first 1/4 HP reciprocating steam engine unit was installed in 1904 by Prince Piero Ginori Conti in the Larderello geothermal field in Italy. Prior to World War II, there were already 136.8 MW of capacity installed in Larderello area. After the war more wells were drilled and modern power stations were installed in the area. As of December 2009, the current operator, ENEL, had 842 MW of installed geothermal power capacity in the Tuscany area.

The first steam engine-driven generator of 35 kW was installed in the USA in 1921 in The Geysers of California. Only in the 1950s, the region was further developed and today 900 MW are produced in this area.

In Japan, surveys began in 1918 with the first experimental generator installed on the island of Kyushu in 1925. Russia followed in 1941 in Kamchatka. Extensive exploration and installations were performed in the 1950s in Japan, Russia, New Zealand, Iceland, Kenya, the Azores Islands off of Portugal and The Philippines.

In the 1980s, Organic Rankine cycle power conversion was applied to geothermal resources of lower enthalpy and widened the range of exploitable resources to lower temperatures.

Today about 10,700 MW are in operation in 24 countries.

Introduction

Power conversion is the least risky part of a geothermal project. Generally it consists of a straight forward engineering design with work executed by experienced manufacturers, engineering firms and contractors.

The risks and challenges are related to exploration, drilling and managing the resource (see preceding entries). Optimization depends on the choice of adaptation of the power station configuration to the resources available (see section on “[Choosing the Energy Conversion Systems](#)”).

There are four basic types of resources:

- Vapor dominated
- Water dominated
- Pressurized water
- EGS engineered geothermal systems where water has to be pumped into the hot rock fissures and cavities. These systems are in early development and demonstration phases

Four energy conversion systems for geothermal resources are in commercial operation:

- Steam Rankine Cycle for Dry Steam
- Steam Rankine Cycle for double or triple flash
- Organic Rankine Cycle in Binary stations for moderate resource temperature
- Combined steam and Organic Rankin Cycle for resource of high temperature and non condensable gases.

To widen the range of resources suitable for power generation beyond dry-steam and flashed steam stations. Of the over 10,000 MW of geothermal stations

installed worldwide, most use steam turbines operating on dry steam or steam produced by single or double flash by the end of 2009. About 1,000 MW use ORC or geothermal combined cycles.

Operational experience has confirmed the advantages of ORC power stations, not only for low-enthalpy water-dominated resources, but also certain high enthalpy ones where the brine is aggressive or the fluid contains a high percentage of noncondensable gas. The higher installation cost of these systems is often justified by environmental and long-term resource management considerations.

This entry is not a design manual for a power station (for detailed calculations, references are given), a history (see preceding entries), or an inventory of existing geothermal stations or a list of future projects. These subjects are dealt with in the following entries.

This is a review of power conversion configuration based on the author's 30 years' experience of implementing various power systems.

What is attempted here is to give a comprehensive picture of adapting the power conversion to various geothermal resources optimizing the output, minimizing the negative implications, and enhancing the sustainability. The objective is to enable the reader to assess the full picture while relying on experienced consultants and vendors in each discipline.

The most updated and comprehensive data on geothermal energy covering both the resource and power conversion systems can be found in: DiPippo R, “Geothermal Power Plants,” 2nd edn. Elsevier and Glassley W, “Geothermal Energy,” CRC Press.

Further reading on geothermal energy can be found on the Department of Energy internet site at:

<http://www.energy.gov/energysources/index.htm>

<http://www.energy.gov/energysources/geothermal.htm>

<http://www1.eere.energy.gov/geothermal/history.html>

The MIT publication on the future of geothermal energy can be found at http://geothermal.inel.gov/publications/future_of_geothermal_energy.pdf

The IEA-GIA Website at <http://www.iea-gia.org/>

The Geothermal Research Council at <http://www.geothermal.org/>

The Geothermal Energy Association at <http://www.geo-energy.org/>

Geothermal Project Design and Implementation

General

A geothermal project is composed of two elements very different in nature and risk: a straightforward conversion system, converting heat into electricity, and the heat supply, i.e., the geothermal resource with attributes similar to oil and gas fields.

Using the terminology of the oil industry a geothermal power project is an integration of “upstream” with “downstream,” hence its particularity. The “upstream” is handled in the other entries. The “downstream” covers the surface equipment.

The project risk is associated with the primary fuel development and it rests with the Investor (Independent Power Producer (IPP) or Utility) rather than with a supplier of fuel.

The aggregate risk in a geothermal project, in a macro sense is different from fossil fuel plant in that the fuel supply risk and investment are mostly up front. As it is site and location specific, it cannot be monetized and the resource supply or quality cannot be substituted, as fossil fuel can, by finding another source from the market.

An important advantage of geothermal power stations is that after the initial investment is made. The power is supplied at a predictable cost unaffected by price fluctuations.

Overview of Geothermal Station Implementation

Development of a geothermal project typically proceeds in two parallel paths, technical operations, i.e., “work on the ground” and commercial/legal procedures, as shown in the scheme in [Table 1](#).

The process starts with identifying a potential site based on desktop studies and preliminary fieldwork covering known geological data and surface manifestations, indicating the existence of an underground geothermal system (hot springs and fumaroles). Results of previous exploration by other parties (other geothermal developers, state agencies, mining companies, oil and gas companies, etc.), if such data exists. This information is then combined with business information, including the need for power in that area, the expected power prices, the existing and missing infrastructure, namely, a transmission grid and road, environmental/

permitting constraints, other business, financial and political risks, etc. A “go/no-go” decision is taken. If a “go” decision is taken, the next steps are as follows:

Obtaining a geothermal concession (when applicable):

Since most countries view geothermal resources as a strategic resource owned by the state, obtaining a state geothermal concession is the first step to geothermal development. This concession is typically awarded by winning a state bid. In the USA, the Federal Steam Act basically assigned geothermal rights to the surface owner, be that the Federal Government Bureau of Land and Management, the state, a private owner, etc.

Obtaining a land position: This is done most typically through geothermal leasing, but sometimes also through land acquisition or other business structures.

Initial nonintrusive exploration: typically including basic geology, geochemical sampling, and geophysical surveys, e.g., gravity, magnetic and electric. This phase may require some permitting work, depending on the type of fieldwork and surveys plans and the local regulations.

Permitting for exploration drilling: Exploration drilling, typically starting with shallow temperature gradient holes, moving to medium depth and slim holes and eventually two to three full diameter deep production and injection wells.

Permitting for roads, power lines, power stations, and its operation.

Long-term multi-well flow testing: In order to accurately determine the average temperature and flow, build a reservoir model and estimate the size of the reservoir and the potential electrical generation.

Technical and economical feasibility and final “go/no-go” decision.

Field testing and initial exploration of the site enables the understanding and establishing of the site potential and drawing station layout, initial heat, mass balance and required investment.

This will allow the commercial and legal part to proceed, typically leading to signing of a long-term Power Purchase/Sales Agreement with a local utility.

Although station design will already start during the period of legal and financial negotiations, the completion of design and manufacturing will not

[illegible]

commence before the resource is appraised and sale of electricity secured by, or for a utility. At this stage, an NTP (Notice to proceed) will be issued to the engineering, procurement, and construction (EPC) contractor and all technical operations released. The steps are as follows:

Station design:

Obtaining all necessary station permits, including the well-field and transmission line.

Completion of well-field development and equipment manufacturing.

While equipment is being manufactured, on-site operations continue including wellhead, piping, interfaces to transmission lines, roads, buildings, power house, control room, etc. They are marked as:

Gathering system and station construction.

Transmission and interaction.

Both actions, completion of manufacturing and on-site construction, end approximately 4 to 5 years from obtaining land position and kick-off of initial exploration. The actual timeframe can be as short as 2–3 years or as long as 10 years, depending primarily on the existence of historical exploration data, the complexity and pace of permitting and logistical constraints.

Power Station Costs

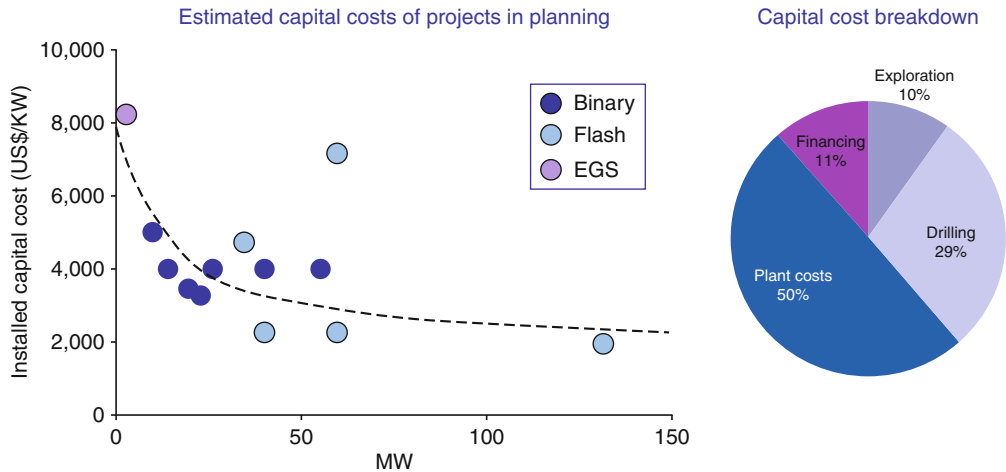
Essentially, the costs of a geothermal power station are highly dependent on the resource. Most of the variation in costs is experienced during the drilling phase where initial exploration holes will define the resource, its temperature, flow rates and chemical and mineral content for which assumptions can be made regarding the cost of managing the resource over time.

Overall costs in 2010 US dollars range between US\$3,000 and US\$5,500 per installed kilowatt depending mainly on the quality of the resource (temperature, geothermal fluid chemistry) which compels the conversion technology (binary, flash, steam). This leads to electricity prices in the range of about US\$70–US\$100 per megawatt-hour.

Cost Estimates for Megawatt-Hour (MWh) Recent estimates from studies on the costs of developing

geothermal power stations provide a range for where actual costs may fall. In September 2008, The US Department of Energy provided a broad range of US \$63–US\$102 per MWh, assuming resource incentives, such as the Federal Production Tax Credit (US DOE 2008 [1]). Several recent estimates put costs within this range, including the California Energy Commission (CEC 2007 [2]), Emerging Energy Research (EER 2009 [3]) and Glacier Partners (Glacier Partners 2009 [4]), which produce estimates between US\$72 and US\$100 per MWh. Further evidence for the cost of power comes from recent contracts approved by the Public Utility Commission of Nevada (PUCN [5]) in July 2010 and released to the public ranged between US\$86 and US\$98 per MWh with resource incentives assumed in the contract price (Source: PUCN 2010 [5]). Internationally, prices may be lower due to economies of scale, but not significantly. The International Energy Agency (IEA) backs this up with a 2010 article that describes costs for new generation is in some countries (such as New Zealand) as highly competitive, ranging from US\$50 to US\$70 per MWh for “known high-temperature resources.” Overall, IEA suggests a wide range of costs, from US\$50 per MWh up to US\$120 per MWh for flash stations. For binary stations, the range in the USA is US\$70 per MWh, up to US\$120 per MWh. However, in Europe where some countries, like Germany, are drilling even deeper for low-moderate temperature resources sufficient for power production, costs may be as high as US\$200 per MWh (IEA 2010 [6]).

Cost Estimates for Installed Kilowatt (kW) As for installed cost per kilowatt, an August 2009 study by the California Energy Commission estimates costs range between US\$2,700 and US\$8,000 for geothermal stations (assuming a station built in 2010), with an average cost assumed of US\$4,851 for binary stations and US\$4,407 for flash stations, although O&M costs are assumed to be higher for flash stations (CEC 2009 [7]). Other recent studies place the range slightly narrower. US DOE, in September 2008, estimated a base cost of US\$4,000 per installed kilowatt. Emerging Energy Research assumes a range of US\$4,000–US\$5,500/kW for projects in the 20–60 MW range (EER 2009 [3]). EER notes that larger projects, “such as those under development in Indonesia, The Philippines, and



Geothermal Power Conversion Technology. Figure 1

Breakdown of capital costs for a geothermal project (Courtesy of IHS Emerging Energy Research [3], pp. 3–18)

New Zealand, have shown significant advantages of increasing scale, with costs of proposed projects approaching as low as US\$2,000/kW” (EER 2009 [3]), refer Fig. 1. According to the International Energy Agency, greenfield flash stations can cost as low as US \$2,000 per installed kW and range up to US\$4,500, particularly in high-temperature sites which may require fewer wells. They estimate binary power stations, ranging from as low as US\$2,400 per installed kW for a productive high-temperature site, suitable for binary technology, to US\$5,900 for low-temperature sites, particularly where many wells need to be drilled (IEA 2010 [6]). Currently, most projects in The Philippines are anticipated to be within the 20–60 MW range, as by an announcement of plans by the Energy Development Company (EDC [8]) on July 29, 2010, where upcoming projects sized 20–50 MW were referenced in its portfolio. EDC estimate costs for greenfield sites (sites that are new and not expansions to producing fields) to cost approximately US\$3,500 per kilowatt installed (Source: Inquirer.net, July 29, 2010).

Comparison with Other Technologies Typically, the cost of the power station, surface facilities and transmission will constitute approximately half of the total costs of a geothermal station (Source: DOE 2008 [1] and EER 2009 [3]). This reflects the high upfront costs associated with resource assessment. For comparison,

the costs to assess solar resources are relatively minor, perhaps even negligible. Yet, materials costs for solar projects are significantly higher on a levelized basis than for fossil fuels or geothermal. Generally more land is needed for economies of scale, which also adds to the costs. For example, in its August 2009 report to the California Energy Commission, KEMA [9] shows that the installed costs for Photovoltaic solar and Parabolic Trough solar technologies are roughly similar in installed cost of geothermal projects overall. However, this recognizes that most of the installed cost is materials, with little of it from actual resource confirmation. Further, capacity factors and annual average production for solar technology are generally much less than for geothermal. Whereas geothermal generally has capacity factors above 90% (CEC 2009 [7]), existing solar technologies have capacity factors between 26% and 29% (CEC 2009 [7] and CPUC 2010 [10]). According to the CEC in its “Renewable Energy Cost of Generation Update,” August 2009 (pp. 206, 211, 226, 236) the capacity factor for geothermal projects averages between 90% and 94% depending on technology. On the other hand, solar thermal without storage and solar photovoltaic is between 26% and 28%. Although solar thermal with storage is expected to have much higher capacity factors (60–70%), no advanced solar projects have proposed capacity factors this high. This is reinforced by the California Public Utilities

Commission (CPUC) which in its July 2010 RPS Project Status table identifies contract capacity for projects in operation and projects which have been approved by the Commission. Solar photovoltaic projects in operation and approved by the CPUC average or expect to average 25.7%, while solar thermal projects approved by the CPUC expect to average 28.8%, refer http://www.cpuc.ca.gov/NR/rdonlyres/A5406F32-B0D0-409E-AA92-0EA79E97BECC/0/RPS_Project_Status_Table_2010_July.xls.

Further, for fossil fuels such as natural gas and coal, the risk of resource development falls to the producers of natural gas and coal, and not to the power station operators. Thus, the risk of resource development is spread widely to all natural gas and coal suppliers, creating a price range (per MMBTU for natural gas and per short-ton for coal). The price of the fuel source is thus reflected in O&M, and not the physical construction of the power station, which varies primarily according to cost of materials like metal and steel.

Should a geothermal resource be more difficult to develop or anticipate, this will be born in initial development cost, rather than the O&M, as the operator of a geothermal power station was also the developer who financed the project and incurred upfront risks. Although geothermal resources vary in sustainability and there is a range of O&M costs, much of the risk is borne in the upfront development process.

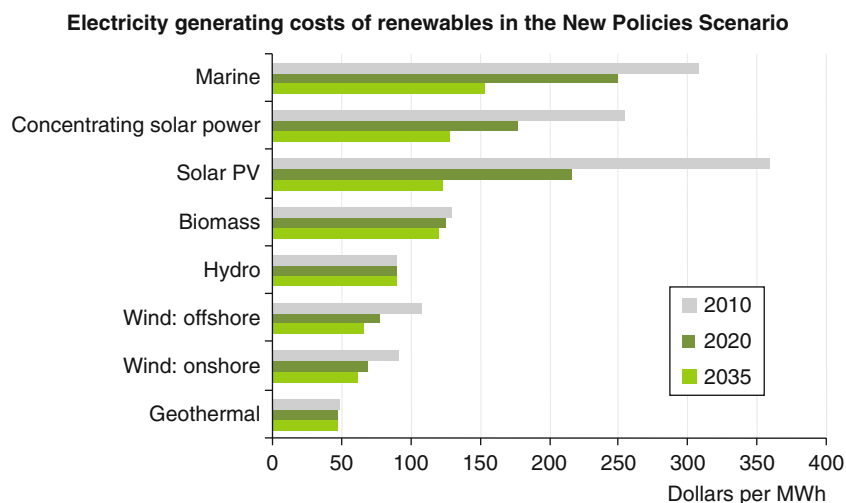
Levelized Costs For the reasons described above, levelized costs for geothermal are important considerations. The levelized cost of energy includes all the costs over the lifetime of a project. From initial investment (higher for geothermal power than fossil fuels), to construction and materials cost (similar for geothermal and fossil fuels), to operations and maintenance costs (slightly lower for geothermal than fossil fuels) and to cost of fuel (essentially zero for geothermal, whereas a major cost for fossil fuels) (Fig. 2).

Geothermal Resources

For power conversion, the value of a geothermal resource is its enthalpy. All other characteristics, mineral content and noncondensable gases, are almost always a drawback to be dealt with at additional investment and cost. In case of coproduction of geothermal energy from pressurized oil or gas wells, the pressure is also of value but its sustainability is in question and requires further R&D.

Geothermal Resource Characteristics

Conventional drilling techniques are used to reach natural underground reservoirs (aquifers) containing hot water and/or steam. These geothermal fluids are under high temperature/pressure and are at bursting pressure from the well. At lower temperatures the



Geothermal Power Conversion Technology. Figure 2
Levelized costs of renewables (IEA World Energy Outlook [6])

geothermal fluid resource has to be pumped to the surface and used to produce electricity via the power conversion system or directly for space or process heating.

While the main aim is to use the heat content for power production, it is essential here to cover all the unwanted source properties (chemical and physical) and the means to mitigate their impact.

Natural geothermal systems can be divided into four categories:

- Dry steam
- Vapor-dominated
- Liquid dominated (superheated water)
- Moderate temperature water (less than 150°C)

In addition, the following are resources in experimental stages:

- Geo-pressured reservoirs
- EGS/HDR
- Lavas and magmas (not dealt with in this entry)

Although dry-steam fields are relatively rare, the Italian fields at Larderello and Mt. Amiata produce about 850 MW [12] and the US field at The Geysers in California, produce about 900 MW (net) of electricity [13, 14]. Vapor-dominated fluids are advantageous for power production as they are usually available at relatively high temperature and pressure. The steam is used to directly drive turbines [15, 16].

The more commonly occurring liquid-dominated systems present a complex utilization problem as reasonably high-pressure vapor must be created for power generation in a conventional turbo-generator unit. The fluid can be partially vaporized by flashing it to a lower pressure, in one or two flashing stages [17]. The vapor then expands in a suitable turbine to produce power [16], or in a binary station using ORC it can heat a secondary fluid vaporizing it at a lower temperature (hydrocarbon). The latter then expands in a turbine, condenses, and is pumped in a continuous closed cycle [16, 17]. A disadvantage of direct steam flashing is that multiple flashing steps are required to attain high conversion efficiencies. Only a fraction of steam is produced with a single-flashing stage. Successive flashing improves efficiency but requires complex turbine design. Furthermore, the relatively high specific volume of steam at these lower temperatures results in

large, expensive turbines. Therefore, in most cases only double-flash is used.

Fluid temperatures as high as 300°C have been observed in the Imperial Valley of California, and Russian fields including Kamchatka. Liquid-dominated systems vary widely in terms of the available temperatures and pressures of geothermal fluids. For power production, temperatures above 150°C are desirable when coupled to sink (heat rejection) temperatures of approximately 25°C.

Geothermal pressured reservoirs such as those of the Gulf Coast, contain moderately hot water (150–180°C) under extremely high pressures (250–650 bars) [18]. However, utilization of this resource has been limited by engineering problems associated with drilling into such formations and extracting useful amounts of energy. Lavas and molten magmas are another potentially useful energy source, but controlled energy extraction is only in the formative research stages at this point, mainly in Iceland.

Enhanced (engineered) geothermal systems (EGS) are also under development in Europe, Japan, Australia, and the Western USA. EGS consists of drilling into hot dry rock (HDR) and creating a geothermal reservoir by hydraulically fracturing the rock [19, 20]. Water is then circulated through the fractured zone to remove heat and is pumped to the surface. Additional surface area for underground heat transfer may also be created by thermal stress cracking which will greatly enhance the lifetime of the reservoir. Energy conversion on the surface may utilize direct steam flashing and steam turbine and/or an Organic Rankine power cycle.

Geothermal Fluid Chemistry

Geothermal fluids contain numerous minerals and dissolved gases accumulated in the underground aquifer from its creation. Utilization of the thermal energy of the extracted geothermal fluid changes the fluid's equilibrium properties when the fluid exits the well-head. Chemical analysis enables the station operator to be prepared for changes in the fluid behavior from changes in temperature and pressure. This parameter is not only required for geothermal power station's (corrosion, scaling, etc.) functional design but also for the station environmental design. Chemical "fingerprints" might adversely affect water quality (domestic,

irrigation, rivers, drainage, etc.). Steam and condensate containing “non-condensable gases” (NCG), e.g., hydrogen sulfide, traces of benzene, toxic (or even only foul-smelling) are handled according to local environmental limitations for such substances.

Reinjection of separated brines, and blow down from cooling towers are common practice in geothermal stations for environmental reasons and for reservoir replenishment.

Noncondensable Gases Two particular ingredients of geothermal fluid production are H_2S and carbon dioxide (CO_2). In most cases, CO_2 outweighs the H_2S concentration and is responsible for relatively low pH of geothermal condensates involved in calcite scaling in production wells. Calcite deposits form more readily in wells with high CO_2 . Scaling is not restricted to well casings and may also impact surface piping and heat exchangers.

Hydrogen Sulfide (H_2S) In certain conditions, H_2S creates a protective iron sulfide layer on carbon steel surfaces (limiting their corrosion rate) and stabilizes “deaeration” of geothermal brines (dissolved oxygen in aerated brines dramatically increases corrosion rates). H_2S is toxic foul smelling, corrosive and scale forming with many heavy metals. A specific type of corrosion (sulfide stress cracking or SSC) can occur on high-strength ferrous materials of the steam turbine blades and/or on “high” hardness welds in geothermal fluids at high H_2S partial pressure. This is hydrogen embrittlement cracking where hydrogen is generated by the sulfide corrosion process on the metal surface.

H_2S readily forms heavy metal sulfides on cooling, or when separating steam from brine. In highly saline geothermal fluids, such sulfides may plug wellheads or heat transfer surfaces. Unique sulfide forms may result in less saline brines after cooling (antimony sulfide “stibnite”). Some heavy-metal sulfide precipitates are also suspected of causing pitting corrosion on stainless steels.

NACE standard MR0175 defines a “sour” (high H_2S containing) service, where H_2S partial pressure is above 0.05 psia (0.0003 Mpa) and requires protection of personnel and electrical equipment.

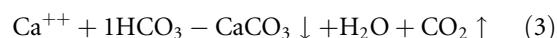
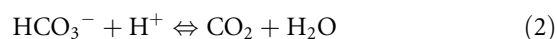
In addition, H_2S is toxic even at low air concentrations and must be mitigated to keep its levels below environmental thresholds.

Calcite Scaling in Production Well Casings One of the common problems in geothermal wells is the deposition of calcium carbonate or calcite, $CaCO_3$, in the well casing starting just above the flash horizon. It is not uncommon for high-temperature geothermal fluids to be close to saturation with respect to calcite as they flow through the formation. The solubility of calcite varies inversely with temperature, so it cannot precipitate from the geothermal fluid because of a decrease in temperature while other factors remain constant. The other properties of the geothermal fluid that influence the solubility are:

- Partial pressure of carbon dioxide CO_2
- pH
- Salinity
- Calcium ion concentration

The first two factors are interrelated. When the geothermal fluid flashes in the well, the released steam carries most of the CO_2 . This causes the liquid pH to rise dramatically supersaturating the geothermal fluid (with respect to calcite). Precipitation occurs immediately and can lead to severe narrowing of the wellbore for several meters just above the flash horizon.

The chemical equilibrium reactions controlling the process are:



pH Geothermal brines and condensates can exhibit different pH values. There are also variations of measurement results related to environmental conditions at the testing point (i.e., difference between “field” and “lab” pH). At higher temperatures, “neutral” pH value lowers while on steam from water separation (flashing to lower pressure) and pH value rises due to the escape of acidic gases.

The impact of the pH value (after dealing with eventual inaccuracies) is important for design purposes:

- Influence on durability of materials in contact with the fluid. Impact on material selection for its transportation and handling (deleterious influence) on the corrosion resistance of carbon steel and lower pH.

- Influence flow rates and pressure drops by indirect effect of scaling impeding heat transfer due to fouling.
- Design may also need the pH value to adjust to lower values (“pH modification” [21]) as an antiscaling measure (to delay amorphous silica precipitation). Value can be raised to protect carbon steel from corrosion in condensate. pH values in some geothermal brines may be highly buffered by a heavy presence of bicarbonate ions. Therefore the input of pH value alone may suffice to adjust for eventual changes throughout the power station.

Dissolved Solids

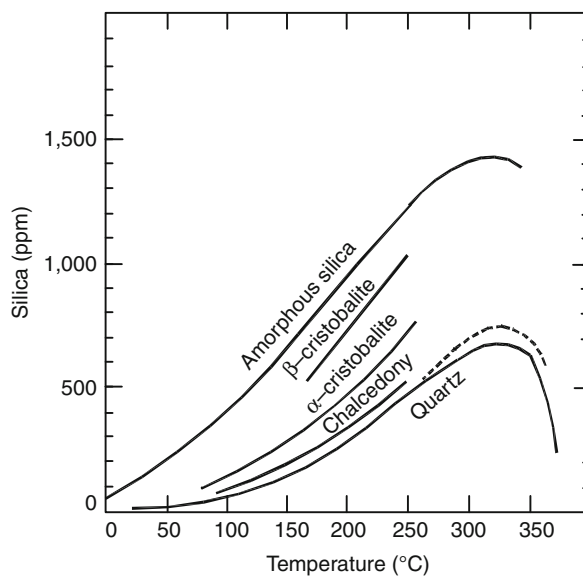
Salinity The importance of this parameter is related to geothermal brine characterization. Residual salinities may also be found in separated steam (from carryover of brine). Salinity impacts heavily on scaling problems, corrosion behavior of steel and other alloys, environmental control, etc.

Power station design includes concentrating brines by flashing the steam from them as well as cooling brines, and then mixing the brines and condensates. All this impacts the salinity and solubility of the solids constituting the brine. Some references [22] consider salinity as dissolved “solids” in water.

At geothermal power extraction pressures and temperatures, solids solubility can either increase continuously with temperature (at saturated water vapor) or decrease [22]. Such behaviors are more complex due to pH changes and gas involvement. Design and further operation have to handle solubility issues to avoid scale plugging (resulting in oversaturation of individual salts and their precipitation). Refer to following paragraph (waste brine scale potential).

Scale in Disposed Geothermal Fluid The common problem discussed in literature [21, pp. 124–128] is the “silica” issue. Here the chemistry is problematic as the resource is at higher temperatures. Silica (SiO_2) solubility in the hot reservoir is controlled by the crystalline quartz form. At lower temperatures, solubility is controlled by amorphous silica typical to waste brine. Other forms of silica may be found in geothermal precipitates as shown in Fig. 3.

The dominant scale precipitate is of amorphous silica which at intermediate temperature possesses



Geothermal Power Conversion Technology. Figure 3 Solubility of various forms of silica in water at saturated water vapor pressures [22, p. 146]

higher solubility than the quartz (brine is “supersaturated” with quartz but undersaturated with amorphous silica).

With proper process design it is possible to avoid precipitation of amorphous silica while the fluid is traveling through the station components. It is definitely possible for precipitation to occur in the injection wells or in the reservoir once the waste fluid returns to the formation. This adversely affects the formation permeability, reducing the injectability of the waste fluid. Any reheating of the waste brine in the formation reduces the potential for precipitation in the reservoir [21, p. 124]. Further aspects of silica precipitation see in [21, pp. 125, 126, 128].

The solubility of silica is not only a function of fluid temperature but also of salinity and pH. The figures shown above are for pure water. Qualitatively, for a given temperature and pH of aqueous solutions, the higher the salinity (i.e., higher molarity), the lower the solubility of both quartz and amorphous silica. For a given temperature and salinity, the solubility of amorphous silica is essentially independent of the pH for low (acidic) values, but rises dramatically as pH climbs above neutral, i.e., $\text{pH} > 7$. The effect is more pronounced for fluids with high salinity.

Precipitation kinetics plays a critical role in the scaling potential in geothermal station components. If precipitation can be slowed it may be possible to process the fluid and dispose of it before scaling can occur. Alternatively, if the precipitation can be accelerated, it may be possible to force the fluid to give up its scale-causing minerals in a rapid and controlled manner before it enters the station proper, allowing the purged fluid to be used without fear of further precipitation. Both of these effects have been used at stations near the Salton Sea in the Imperial Valley of the USA where highly mineralized, corrosive brines are present. There are five parameters that influence the kinetics of the silica precipitation (essentially a polymerization process):

- Initial degree of super saturation (i.e., actual SiO_2 concentration – equilibrium concentration)
- Temperature
- Salinity or molarity of the solution
- pH of the solution
- Presence (or absence) of colloidal or particulate siliceous material

The first and second factors can be controlled by proper selection of separator and flash conditions for a given geothermal fluid. The third factor is a fluid characteristic that cannot be controlled. The fourth and fifth factors can be adjusted as the fluid moves through the station from the production wells, through pipes, other components and eventually to the injection wells. When the brine is acidified, the rate of precipitation is very slow and the fluid can be viewed as temporarily stabilized. As the pH is raised, the precipitation rate increases dramatically, reaching a maximum at near-neutral pH values, about 6.0–7.5, and then slowing as pH approaches 9.0–9.5. The rates for pH = 5.3 and pH = 9.0 are roughly the same.

The last factor in the list has been utilized successfully for the Imperial Valley stations. Geothermal fluid is “seeded” with silica particles in large vessels called flash-crystallizers. These provide favorable precipitation sites for the supersaturated solution. After two stages of this process, the precipitated silica is removed, dried, compacted and disposed of. The generated steam is ready for use in turbines and the waste liquid is sufficiently clean for reinjection without fear of clogging the reservoir.

The potential for silica precipitation is mitigated to some degree when binary stations are used as the geothermal fluid is not flashed, but only cooled. Thus, there is no increase in the concentration of silica as the fluid passes through the station. Flow design in binary stations keeps the fluid in the safe region below the amorphous silica equilibrium curve. In comparison to a flash station, this allows the geothermal fluid to be cooled to a lower temperature before silica precipitation occurs.

Entrained Solids Entrained solids consist of sand and clay particles requiring temporary or permanent filters (stand or centrifugal type) to avoid corrosion damage to pipes and heat exchangers as well as clogging of the injection wells.

Material Selection in Geothermal Power Stations This issue was presented and discussed by Kestin in the Source Book on the *Production of Electricity from Geothermal Energy*-Chap. 3 [16], See Table 2.

Different applications within the power station face different requirements which are not only related to mechanical design, but also to durability problems arising from the contact with the complex geothermal environment.

Metallic Materials Most of the construction materials used in such stations are metallic. Performance of metals and alloys in geothermal power stations has

Geothermal Power Conversion Technology. Table 2
Typical turbine element materials [16]

Component	Material
Piping	ASTM A106, Gr B; ASTM A335, GrP11 or P22
H.P. castings	ASTM A356, k Gr 1, 6, 9 or 10
L.P. castings	ASTM A285 or A515
Valve bodies	ASTM A216 or A217
Fasteners	ASTM A193 and A194
Rotors	ASTM A470
Blades	AISI 403
Nozzle blades	AISI 403
Bands	AISI 405

long been studied in respect to durability against availability, ease, and cost of fabrication (as in the use of steel and iron base alloys).

Furthermore, mechanical design may reach a conflict with special types of corrosion (sulfide stress cracking) that might occur during the use of some equipment made of high-strength ferrous alloys.

Different types of corrosion and environmentally assisted cracking mostly dictate the materials selection in geothermal power stations. The following corrosion and cracking forms can appear:

- Generalized (or “uniform”) corrosion
- Localized corrosion (pitting, “crevice”)
- Stress corrosion cracking (mostly induced by chloride ions)
- Sulfide stress cracking
- Fatigue corrosion (synergistic effect of cyclic stress and corrosion)
- Erosion-corrosion (accelerated corrosion by impingement of particles, gas bubbles, droplets, or too high fluid velocity)

Poor operating conditions might impact durability, even after applying a good design and materials selection. Among these conditions:

- “Shutdown corrosion,” mostly generating or accelerating uniform and localized attack due to air (oxygen) ingress, water stagnation, corrosion products transformation. This corrosion is caused by frequent or long-term shutdowns without proper preservation actions.
- Bad monitoring or no application of designed corrosion inhibitors added to geothermal fluid.
- Bad or no monitoring of the designed steam quality.
- Unsuitable locally made repairs (like welding).
- Introduction of new geothermal fluid resources.
- Change of flow and temperature conditions.

Geothermal steam turbines are normally fed directly from the production wells, after separation from brine or as slightly superheated steam. Corrosive gases and entrained salts (from brine carryover) impose serious design problems on the turbine, steam transportation system and other ancillary parts. The situation may be more complex due to variations in steam composition (even at the same station).

Geothermal steam turbine manufacturers have established standards for materials selection based on manufacturer experience and material testing:

- For turbine rotors (forgings), some [24] used “low chrome-moly” steel (CrMo steel) and others [16] used “chrome-moly-vanadium” steel or “nickel-chrome-moly-vanadium” steel.
- For moving blades 13 chromium stainless steel.
- For the casings, grey cast iron or carbon steel.
- For steam pipes (large diameter) and silencers, rolled steel.

Sometimes cladding or coating is applied on critical parts:

- Epoxy coating on carbon steel exhaust duct and discharge pipes.
- “Stellite” hard facing on governing valve disc made of cast steel or on check valve made of forged steel [25].
- Austenitic stainless steel type 316L (S.S. 316L) was used as cladding on carbon steel used for condenser shell and for steam separator casing.

Steam turbine equipment also consists of other related parts in contact with steam, separated wet gases and wet air which comprise (partial list):

- Ejectors exhaust and tail pipes (S.S. 316L), condenser spray nozzles (S.S. 316L)

In binary geothermal power stations that are using organic fluids (propane, butane or pentane) there is no corrosion on the motive fluid side. This enables the choice of materials based on mechanical strength only, (i.e., low alloy carbon steel, aluminum or titanium for high speed centripetal expanders) such as low alloy carbon steel, aluminum or titanium for high speed centripetal expanders. The main corrosion issue is encountered in the piping and tubing parts in contact with the geothermal fluid. In most such applications, the use of mild steel benefits from the total deaeration of the geothermal brines which naturally occurs at normal pressurized fluid conditions. Separated brines also benefit from the loss of corrosive CO₂ and H₂S, (of which a high proportion escapes with the separated steam), and of the resulting higher pH. Both factors (oxygen and pH) may contribute to lower corrosion rates and to

reasonable durability of such low-cost material. In some cases, high salinities or high sand content may generate localized corrosion, erosion-corrosion or under-deposit corrosion, requiring different materials such as stainless steel (including duplex) or titanium.

Due to the high presence of corrosive CO₂ in such a fluid, special precautions must be applied on transporting downhole pumped brines inside mild steel piping.

When thin wall tubing (heat exchangers used to evaporate or preheat the working fluid) is applied, stainless steel or even Titanium is selected. Recently, “duplex” stainless steel tubes have been successfully used [23].

Atmospheric Corrosion Geothermal power stations’ atmospheric environment might also have an abnormal presence of H₂S and NH₃ gases. These gases together with humidity can attack copper and other metallic materials. This is in addition to carbon steel rusting due to CO₂ presence with steam escaping into the atmosphere.

Galvanized steel, stainless steel, aluminum, and epoxy coatings are preferably used against atmospheric corrosion. Electronic equipment requires special attention, i.e., pressurized shelters, corrosion resistant contacts, and appropriate insulation.

Nonmetallic Materials While nonmetallic materials are not useful in heat transfer equipment (used only for sealing purposes), such materials find use in the transportation of cooled brines or cooling water pipes (made of fiber reinforced plastics (FRP) or polyethylene).

Internally cement-coated steel pipes are used to transport hyper saline geothermal brines. Cements also find large applications in geothermal well completion and civil works.

Thermodynamic Analysis of the Energy Conversion Process

Introduction

For a comprehensive thermodynamic analysis including Energy analysis applied to geothermal power

systems from fluid supply to energy conversion including ancillary equipment (separators, cooling towers, etc.) [33].

Available Energy

Energy Conversion Process in Geothermal Power Stations It is assumed that geothermal fluids are utilized in continuous steady state in the energy conversion process. This assumption allows thermodynamic analyses of such systems whether water or steam dominated. The First and Second Law of thermodynamics are used to calculate the maximum available energy and the expected cycle efficiency.

First and Second Law of Thermodynamics The First Law of thermodynamics is the application of the conservation of energy principle to heat and thermodynamic processes.

Maximum Mechanical Energy Available Kestin [26] converting heat into mechanical energy is governed by the Second Law. The maximum mechanical energy which can be obtained from the geothermal fluid provided by a constant temperature (T_H) reservoir is achieved by using an ideal heat engine (the Carnot engine). This produces work and discards heat into a low constant temperature (T_L) reservoir.

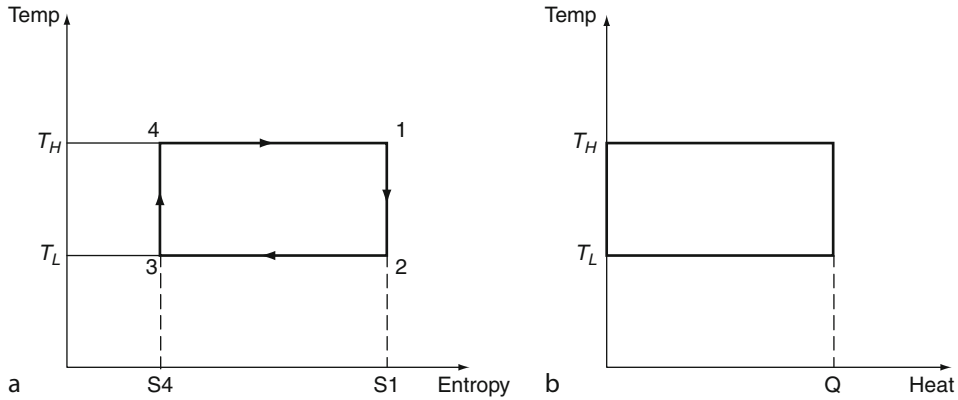
The maximum conversion efficiency is expressed by:

$$\eta_C = \frac{T_H - T_L}{T_H} \quad (4)$$

T-Q diagrams are used to allow direct energy analysis of the heat source and heat utilization. A T-Q diagram of the general Carnot cycle of Fig. 4a is given in Fig. 4b. The above is applied to the four different categories of fluids obtained from geothermal sources where T_H and T_L are the absolute temperatures of the heat source and heat sink.

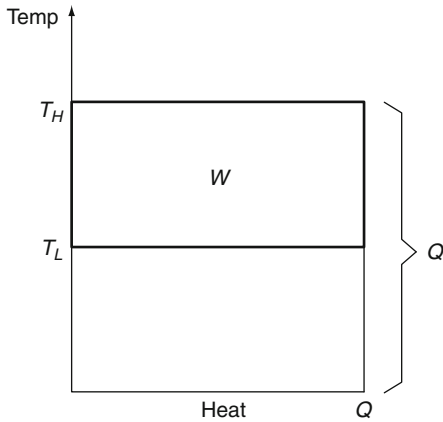
Dry Steam The supply of steam can be considered a constant temperature heat source where the ambient can be considered a constant temperature heat sink.

The relevant T-Q diagram is given in Fig. 5 which is similar to that of Fig. 4b.



Geothermal Power Conversion Technology. Figure 4

T-S (a) and T-Q (b) diagrams showing the available mechanical energy



Geothermal Power Conversion Technology. Figure 5

T-Q diagram of dry-steam utilization

$$\eta_C = \frac{T_H - T_L}{T_H} \quad (5)$$

Where Q is the heat source and W , the corresponding work obtained:

$$W = \frac{T_H - T_L}{T_H} Q \quad (6)$$

Pressurized Water In this case, the temperature of the heat being transferred to the cycle by the geothermal water drops during the heat transfer operation and therefore Eq. 5 does not apply here. To calculate maximum efficiency the cycle is represented by a series of infinitesimally narrow Carnot cycles between temperatures T_H and T_L , see Fig. 6a.

Therefore for each engine i :

$$dW_i = \frac{T_i - T_L}{T_i} dQ \quad (7)$$

and for the cycle:

$$W = \int_0^Q \frac{T - T_L}{T} dQ = \left[1 - \frac{1}{\frac{T_H}{T_L} - 1} \right] Q \quad (8)$$

The efficiency is:

$$\eta = 1 - \frac{1}{\frac{T_H}{T_L} - 1} \quad (9)$$

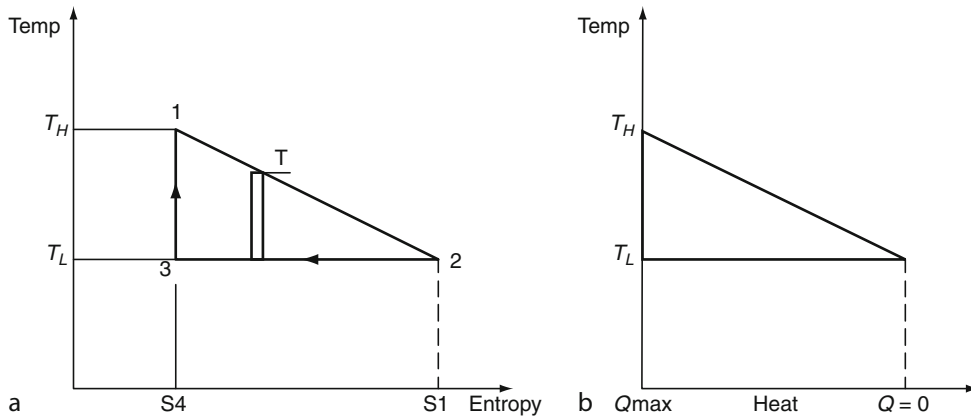
Steam Dominated:

1. “Dry Steam” (majority of energy in the steam)
2. “Pressurized Water” (majority of energy in the water or sensible heat part) (Fig. 7).

If Q is the total heat constant of the fluid and r is the ratio of the latent heat portion of the total heat. Then the maximum efficiency of mechanical energy available from this source is:

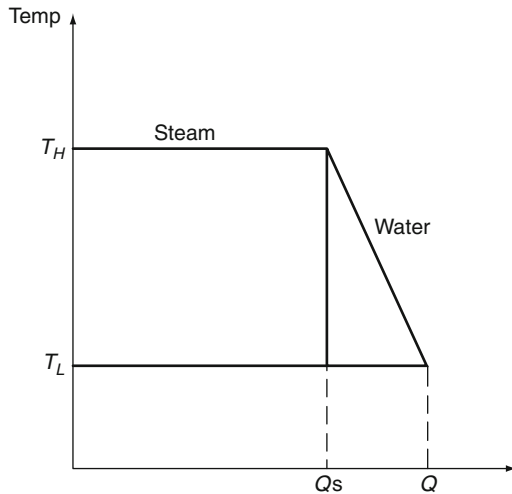
$$W = \left\{ \left(\frac{T_H - T_L}{T_H} \right) r + (1 - r) \left[\frac{1 - \ln \frac{T_H}{T_L}}{\frac{T_H}{T_L} - 1} \right] \right\} Q \quad (10)$$

Liquid Dominated Liquid-dominated case is as in the section on “[Steam Dominated](#)” with most of the energy in the water or brine portion. The formula here is the same but the maximum available mechanical energy



Geothermal Power Conversion Technology. Figure 6

T-S (a) and T-Q (b) diagrams for pressurized or low-temperature water cycle



Geothermal Power Conversion Technology. Figure 7

T-Q diagram for steam-dominated cycle

(or efficiency) from the same total heat is smaller due to smaller r (Fig. 8). In both cases above the combined efficiency depends on r and is expressed by:

$$Ws = w \left\{ \left(\frac{T_H - T_L}{T_H} \right) r + (1 - r) \left[\frac{1 - \ln \frac{T_H}{T_L}}{\frac{T_H}{T_L} - 1} \right] \right\} Q \quad (11)$$

Maximum Specific Power Using Eq. 11, the maximum specific energy (kW) contained in a unit mass

flow (kg/s) of a given heat source, which corresponds to maximum specific power (kW s/kg) is:

$$Ws_{\max} = \dot{W}_{\max} / \dot{m} = h_1 - h_0 - T_0(s_1 - s_0) \quad (12)$$

In Eq. 12, state 1 corresponds to the fluid high-temperature condition and state 0 corresponds to the ambient or heat sink condition.

For a given fluid, a fixed T_0 and reinjection temperature of the geothermal fluid $T_{\text{gf}}^{\text{out}}$, having a minimum value of T_0 , the maximum work (per unit weight of geothermal fluid) possible from an ideal, reversible process is a function of $T_{\text{gf}}^{\text{int}}$ only, the geothermal source temperature. Calculations of the relevant curves are given Fig. 9 for both saturated geothermal steam and water. Any real process will have inefficiencies or nonreversible steps that will result in net work less than Ws .

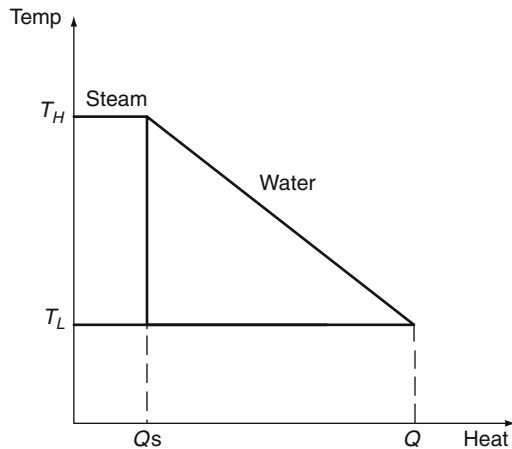
The curves in Fig. 9 allow quick assessment of the maximum power output that can be achieved from a geothermal well producing either water or steam or both. The required parameters are the fluid mass flow rate (in kg/s), fluid mass equilibrium initial temperature, and the ambient temperature or cooling water temperature.

The Temperature Limitation

- The maximum temperature at the inlet to the turbine is T_{Hb} , which is the temperature at the well head.

In practice, the inlet temperature to the turbine is lower because of either the flashing process or the heat transfer in the evaporator.

- The lowest possible temperature T_L of the heat sink (the temperature of condensation at the turbine exhaust) is the ambient temperature T_0 .
 - In practice this temperature is higher because of the irreversibility condensers, evaporator, pre-heaters, and cooling towers.



Geothermal Power Conversion Technology. Figure 8
T-QS diagram for liquid-dominated cycle

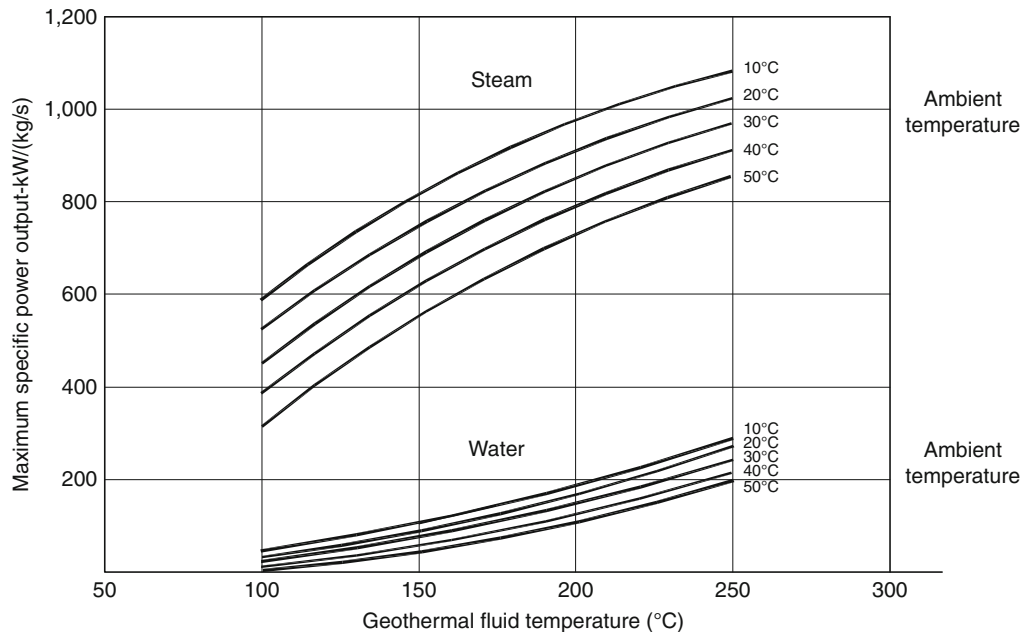
Power Conversion Processes Cycles

Steam Rankine Cycle Dry-steam power stations were the first type of geothermal power stations to achieve commercial status. The first small steam engine was operated in 1904 at Larderello in the Tuscany region of Italy [12].

Dry-steam power stations are simpler and less expensive than flash-steam or binary power stations as there is no geothermal brine to deal with.

Large dry-steam reservoirs have been discovered only in two areas of the world, Larderello and The Geysers. There are limited dry-steam areas in Japan (Matsukawa), Indonesia (Kamojang), New Zealand (Poihipi Road section of Wairakei) and the USA (Cove Fort, Utah). White [27] estimated that only about 5% of all hydrothermal systems with temperatures greater than 200°C are of the dry-steam type.

The general characteristic of a dry-steam reservoir is that it comprises of porous rocks featuring fissures or fractures, either occluded or interconnected, that are filled with steam. Whereas the steam also contains gases such as carbon dioxide, hydrogen sulfide, methane and others in trace amounts, there is little or no liquid present.



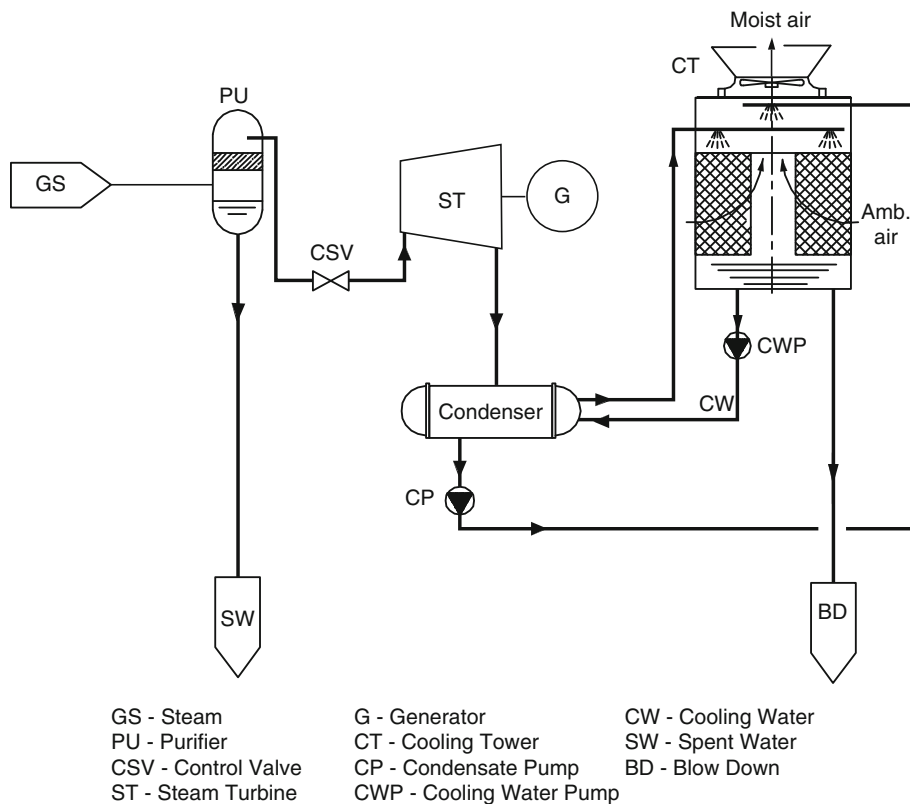
Geothermal Power Conversion Technology. Figure 9

Maximum specific power output as function of geothermal fluid (steam and water) temperature and ambient temperature

The dry steam extracted from the mentioned resources is either saturated or slightly superheated at temperatures near 235°C and at a pressure near the maximum saturation in Mollier curve (30.7 bar). Isenthalpic pressure loss in the upper layers explains the superheated condition at the turbine inlet. Steam from dry-steam reservoirs is superheated due to the pressure drop during the flow through the hot rock at constant temperatures. James [28] estimated the superheating of up to 35°C (above the saturation point). Typical dry-steam power conversion system is given in Fig. 10. The steam only needs final purifying before being sent to the turbine with the condensate used as makeup for the cooling tower. The wastewater collected at the bottom of the purifier and the blow-down of the cooling tower are injected to the aquifer. This eliminates environmental problems and helps in renewal of the aquifer liquids balance.

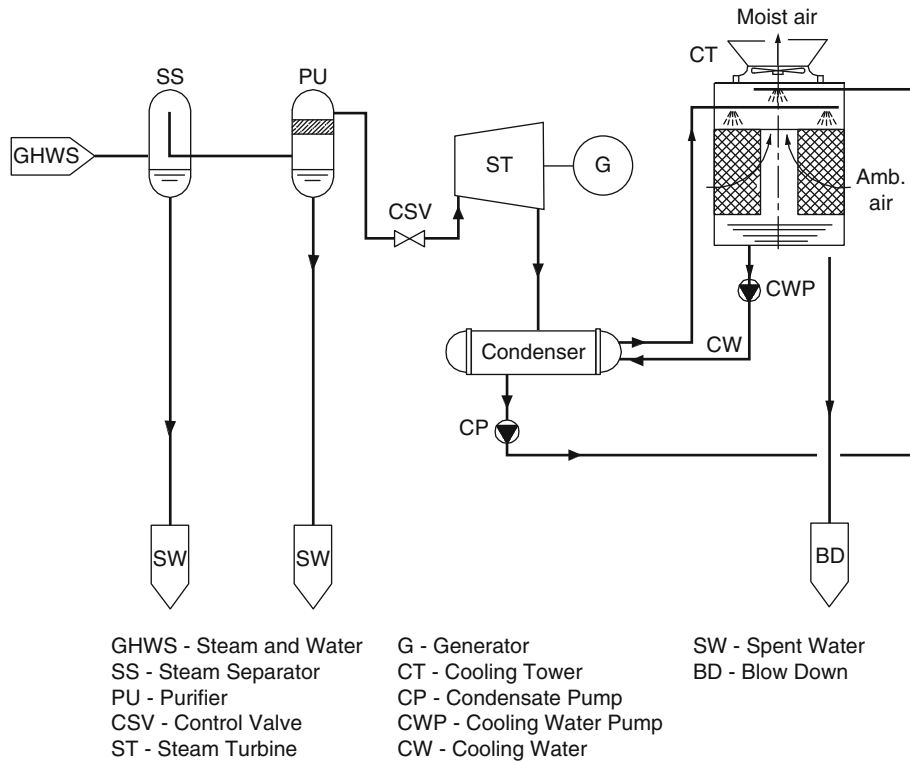
The power conversion section of a single-flash system is similar to that of dry steam except for the steam preparation that requires adequate separation between the steam and water ahead of the steam purifier as given Fig. 11.

Expansion Process Thermodynamic state diagrams are used for easy understanding of the fluid working cycle. A temperature–entropy (T-S) diagram for the single-flash station is shown in Fig. 12. The Mollier h-s diagram that is alongside it may be preferred as the vertical axis shows the exact turbine work $h_4 - h_5$, and it is easily used for efficiency evaluation. The turbine efficiency in the wet zone is lower than in the dry zone mainly due to appearance of small droplets in the expanding steam. To find the turbine efficiency, use a Mollier diagram as in Fig. 12 where $\Delta h_{is} = h_4 - h_{5s}$. Make a first assumption of 0.8 which finds X_{51} ,

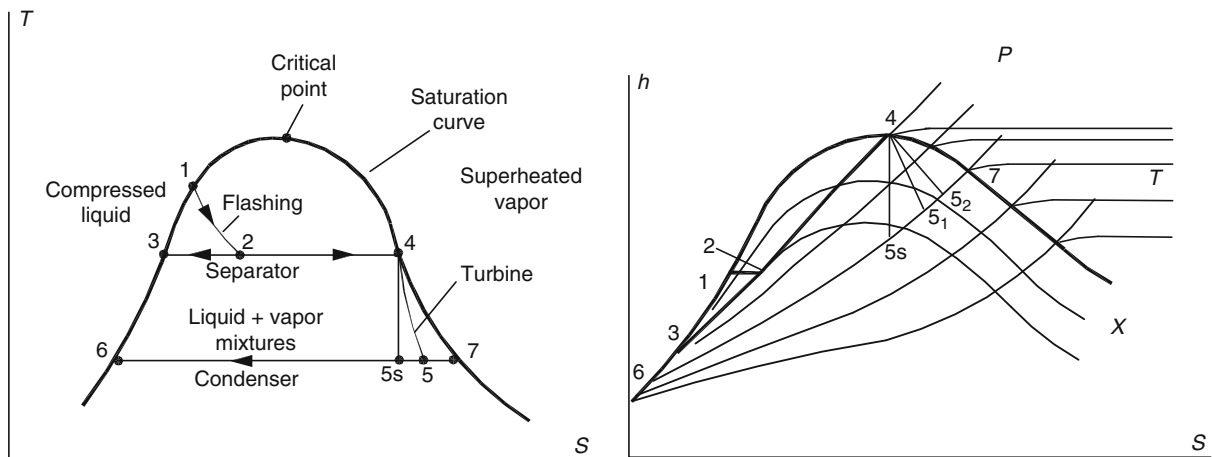


Geothermal Power Conversion Technology. Figure 10

Power conversion system for dry steam



Geothermal Power Conversion Technology. Figure 11
 Power conversion system for single-flash steam supply



Geothermal Power Conversion Technology. Figure 12
 T-S and H-S state diagrams for single-flash stations

(steam dryness at state point 51. The dryness fraction is $x = \text{Mass of dry steam} / \text{Total mass of wet steam}$). According to this the turbine efficiency can be found

by use of the information on turbine efficiency in the dry and wet zones according to Bauman in [29] or Thermoflow practical assumptions [30].

$$\Delta h_{\text{first}} = h_4 - h_{51} = 0.8 \Delta h_{\text{is}} \rightarrow X_{51} \rightarrow \eta_{\text{isw}} = \eta_t \quad (13)$$

From η_t obtain the final enthalpy drop:

$$\Delta h_{\text{final}} = h_4 - h_{52} = \Delta h_{\text{is}}^* \eta_t \quad (14)$$

The steam process of a fuel-driven dry-steam station with and without superheating is shown in Fig. 12. Water is pumped to the boiler (1–2), heated (2–3–4), steam is produced that expands from point 4 to 6 in the non-superheated or from 4 to 6 in the superheated case. At the lower pressure the steam is condensed and pumped back to the boiler to complete the cycle (Fig. 13).

In the geothermal dry-steam case the cycle is partial. Since the wells produce saturated steam (or slightly superheated steam), the starting point is located on the saturated vapor curve. The turbine expansion process 1–2 generates less power output than the ideal, isentropic process 1–2s. Heat is rejected to the surroundings in the condenser via the cooling water in process 2–3.

The actual work produced by the turbine per unit mass of steam flowing through it is given by:

$$w_1 = h_1 - h_2 \quad (15)$$

The maximum possible work would be generated if the turbine operated adiabatically and reversibly, i.e., at constant entropy or isentropically. Therefore, the

isentropic turbine efficiency η_t , is the ratio of the actual work to the isentropic work, namely:

$$\eta_t = \frac{h_1 - h_2}{h_1 - h_{2s}} \quad (16)$$

The power developed by the turbine is given by:

$$\dot{W}_1 = \dot{m}_s w_1 = \dot{m}_s (h_1 - h_2) = \dot{m}_s \eta_t (h_1 - h_{2s}) \quad (17)$$

The gross electrical power will be equal to the turbine power multiplied by the generator efficiency:

$$\dot{W}_e = \eta_g \dot{W}_1 \quad (18)$$

The net power is further reduced by all parasitic loads including condensate pumping power, cooling tower fan power, etc.

High-Pressure and Low-Pressure Turbine Expansion Processes The processes for the double-flash turbine are shown in Figs. 14 and 15 below. The second flash is of the brine separated by the first steam separator.

Using the assumptions of no heat losses, no change in potential and kinetic energy as in a single-flash system, (refer to “Expansion Process”), the power generated by each separated stage of the turbine can be evaluated.

The expansion work of the HP-stage is expressed by:

$$w_{\text{hpt}} = h_4 - h_5 \quad (19)$$

As in the case of the single flash the efficiency is:

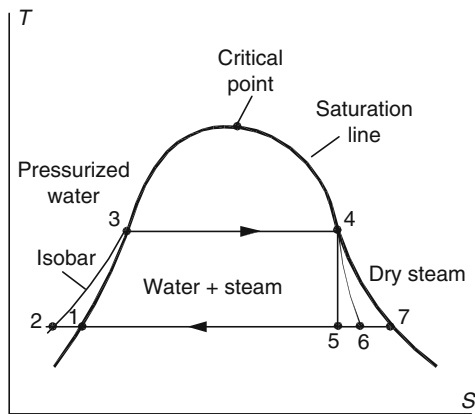
$$\eta_{\text{hpt}} = \frac{h_4 - h_5}{h_4 - h_{5s}} \quad (20)$$

Accordingly, the total work depends on the mass flow:

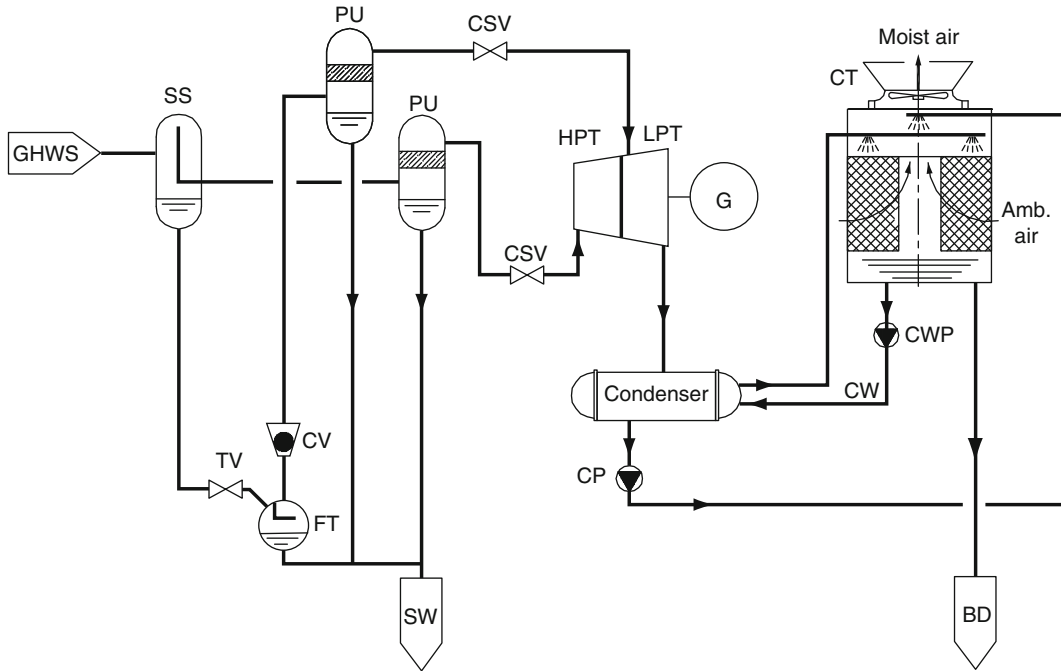
$$\dot{W}_{\text{hpt}} = \dot{m}_{\text{hps}} w_{\text{hpt}} = x_2 \dot{m}_{\text{total}} w_{\text{hpt}} \quad (21)$$

X_2 is dryness at state 2 (see Fig. 15). The actual outlet state from the high-pressure section of the turbine is found using the Baumann rule [29] and by use of a Mollier $h-s$ steam chart [31, 32].

$$h_5 = \frac{h_4 - A \left[1 - \frac{h_6}{h_7 - h_6} \right]}{1 + \frac{A}{h_7 - h_6}} \quad (22)$$



Geothermal Power Conversion Technology. Figure 13 Temperature-entropy diagram for dry-steam power station (steam saturated at the turbine inlet)



GHWS - Steam and Water
 SS - Steam Separator
 PU - Purifier
 CSV - Control Steam Valve
 HPT - High Pressure Turbine
 LPT - Low Pressure Turbine

G - Generator
 CT - Cooling Tower
 CP - Condensate Pump
 CWP - Cooling Water Pump
 CW - Cooling Water

SW - Spent Water
 BD - Blow Down
 CV - Check Valve
 FT - Flash Tank

Geothermal Power Conversion Technology. Figure 14
 Double-flash system with dual-pressure turbine

where the factor A is defined as:

$$A = 0.425(h_4 - h_5) \quad (23)$$

The low-pressure steam from the flasher is actually saturated vapor (state 8) which is admitted to the steam path and joins the partially expanded high-pressure steam at state 5. The mixed steam, ready to enter the low-pressure turbine stages is at state 9. The First Law of thermodynamics and conservation of mass allow finding the properties of the mixed state 9:

$$\dot{m}_5 h_5 + \dot{m}_8 h_8 = (\dot{m}_5 + \dot{m}_8) h_9 \quad (24)$$

$$h_9 = \frac{x_2 h_5 + (1 - x_2) x_6 h_8}{x_2 + (1 - x_2) x_6} \quad (25)$$

The low-pressure turbine may now be analyzed as follows:

$$w_{lpt} = h_9 - h_{10} \quad (26)$$

$$\dot{W}_{lpt} = \dot{m}_9 (h_9 - h_{10}) = (\dot{m}_5 + \dot{m}_8) (h_9 - h_{10}) \quad (27)$$

Again, using the Baumann rule [29] and steam Molier chart [31, 32]:

$$h_{10} = \frac{h_9 - A \left[x_9 - \frac{h_{11}}{h_{12} - h_{11}} \right]}{1 + \frac{A}{h_{12} - h_{11}}} \quad (28)$$

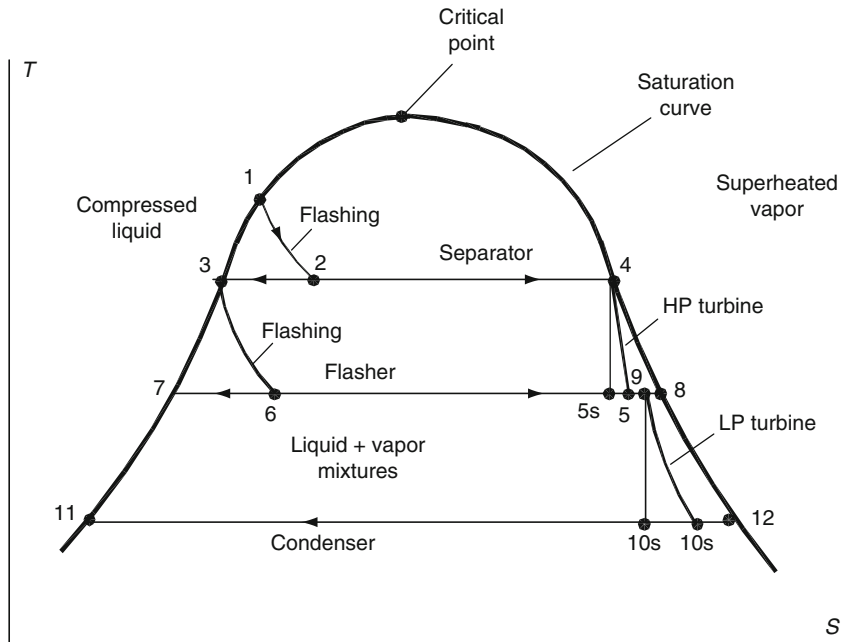
$$A = 0.425(h_9 - h_{10s}) \quad (29)$$

$$h_{10s} = h_{11} + [h_{12} - h_{11}] x \left[\frac{s_9 - s_{11}}{s_{12} - s_{11}} \right] \quad (30)$$

$$\eta_{lpt} = \frac{h_9 - h_{10}}{h_9 - h_{10s}} \quad (31)$$

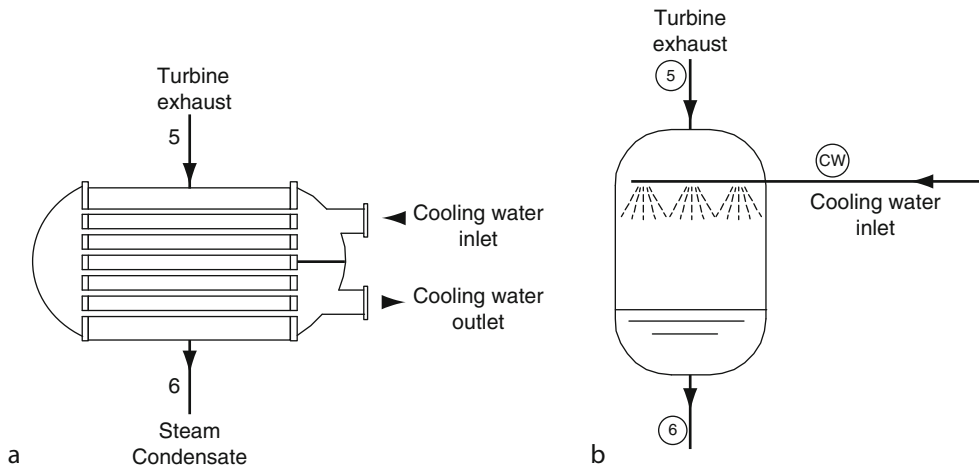
The total power generated is the sum of the power from each turbine:

$$\dot{W}_{total} = \dot{W}_{hpt} + \dot{W}_{lpt} \quad (32)$$



Geothermal Power Conversion Technology. Figure 15

Temperature-entropy process diagram for double-flash station with a dual admission turbine



Geothermal Power Conversion Technology. Figure 16

Surface (a) and direct contact (b) condensers

Finally, the gross electrical power is found from:

$$\dot{W}_{e, \text{gross}} = \eta_g \dot{W}_{\text{total}} \quad (33)$$

Condensing Process In the use of surface-type condenser shown in Fig. 16a, the required flow rate of

cooling water \dot{m}_{cw} related to the steam flow rate $X_2 \dot{m}_{\text{st}}$ is expressed by the First Law of thermodynamics as:

$$\dot{m}_{\text{cw}} = X_2 \dot{m}_{\text{st}} \left[\frac{h_5 - h_6}{\bar{c} \Delta T} \right] \quad (34)$$

where \bar{c} is the assumed constant specific heat of the cooling water (4.2 kJ/kg.K), ΔT is the rise in cooling water temperature at the condenser inlet and outlet and X_5 is the steam dryness stage at the turbine exit.

For a direct-contact condenser (Fig. 16), the equation is:

$$\dot{m}_{cw} = x_2 \dot{m}_{total} \left[\frac{h_5 - h_6}{\bar{c}(T_6 - T_{cw})} \right] \quad (35)$$

Overall Thermal Efficiency The performance of the entire station may be assessed using the Second Law of thermodynamics. This by comparing the actual power output to the maximum theoretical power that could be produced from the given geothermal fluid. This involves determining the rate of energy carried into the station with the incoming geothermal fluid.

The specific maximum energy of a fluid that has a pressure, P , temperature T in the presence of an ambient pressure P_0 and an ambient temperature T_0 , is given by:

$$\dot{W}_{max} = h(T, P) - h(T_0, P_0) - T_0[s(T, P) - s(T_0, P_0)] \quad (36)$$

To get the maximum theoretical thermodynamic power, this term is multiplied by the total incoming geothermal fluid mass flow rate:

$$\dot{E} = \dot{m}_{total} \cdot \dot{W}_{max} \quad (37)$$

The ratio of the actual net power to the maximum power is defined [33] as the utilization efficiency or the Second Law efficiency of the station:

$$\eta_u \equiv \frac{\dot{W}_{net}}{\dot{E}} \quad (38)$$

Stations can be designed to maximize η_u when the value of the primary energy is a significant factor in the economics of the operation.

Organic Rankine Cycle Configurations For low-temperature resources, the efficiency of the flash steam cycle led many researchers to propose cycles which enhance the thermal efficiency and utilization of the geofluid energy content. Some of the ideas were executed into working power stations but did not mature into a commercial stage. The systems that are commercial begin with the simple Organic Rankine cycle (using

pentane), or super critical cycles (using butane or fluoro-carbons), geothermal combined cycle, and other additional similar concepts that will be discussed hereafter. The noncommercial, experimental ideas is included in section on “Experimental Power Stations.”

Ideal Organic Rankine Cycle As already mentioned in “Pressurized Water,” the hot water or brine is not an isothermal heat source as it cools while transferring heat to the working fluid. A more realistic ideal cycle for a geothermal binary station is a triangular cycle consisting of an isobaric (constant pressure) heat addition process up to the brine inlet temperature T_H , followed by an isentropic expansion and an isothermal heat rejection process at T_L to complete the cycle. See Fig. 17.

The efficiency for such a cycle was expressed in Eq. 9 as:

$$\eta_{TC} = 1 - \frac{T_L}{T_H - T_L} \ln \left(\frac{T_H}{T_L} \right) \quad (39)$$

For the same temperature range of 150°C and 40°C, the triangle cycle yields an efficiency of 14.3% compared with 26% for the constant temperature case.

Organic Rankine Cycle Based Power Generation Process Unlike dry-steam and flash-steam power stations, binary stations do not have condensate to serve as makeup for a water cooling tower. As a result, binary stations need a separate cooling medium, either fresh water or air.

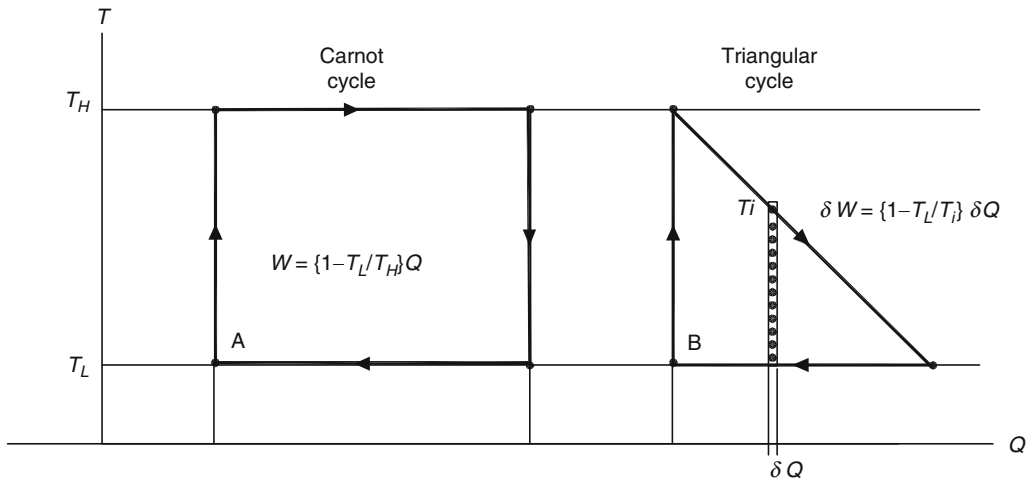
In its simplest form, the binary station follows the schematic flow diagram in Fig. 18 for a water-cooled system [35] and in Fig. 19 for an air-cooled system [36].

The working fluids thermodynamic process are shown in Fig. 20. Due to the inclination of the saturation curve, the vapor expansion extends further into the superheated zone.

The main cycle components will be analyzed later using the state points on Fig. 20.

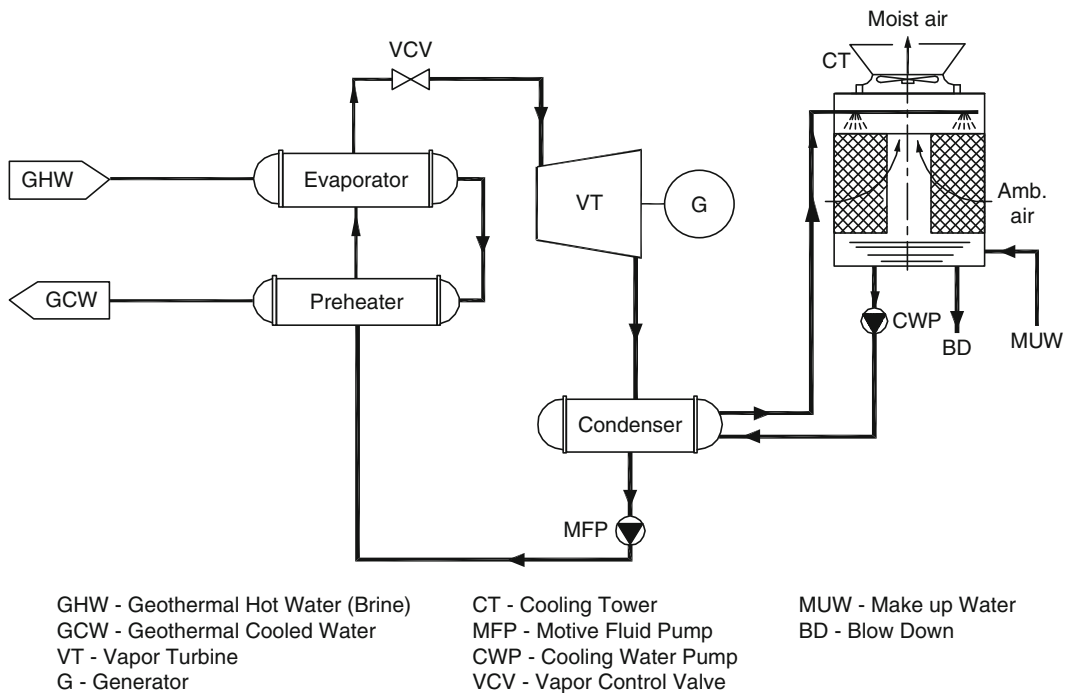
The supply of heat to the Organic Rankine cycle completes the conversion process and is shown in the two cases. The supply of hot water/brine or supply of steam as shown in Fig. 21. In (a) the sensible heat supply is only by hot water or brine, while in (b) the sensible heat and latent heat are from steam.

Hot geofluid is supplied to the evaporator and from there it flows to the preheater and is then returned to



Geothermal Power Conversion Technology. Figure 17

Two ideal thermodynamic cycles: constant temperature and continuous reducing temperature



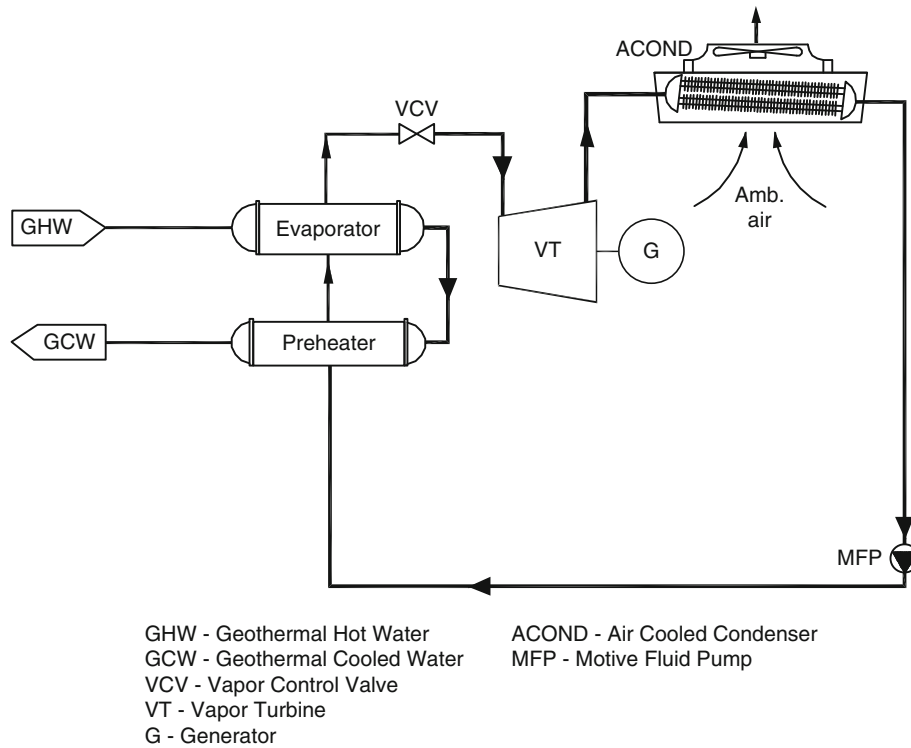
Geothermal Power Conversion Technology. Figure 18

Simplified schematic of a water-cooled binary geothermal power station [35]

the injection well. The working fluid flows through a preheater where it is brought close to its boiling point. It then flows to the evaporator E where it acquires the supplement heat of evaporation. Emerging as a saturated vapor it expands in the turbine, condenses in the

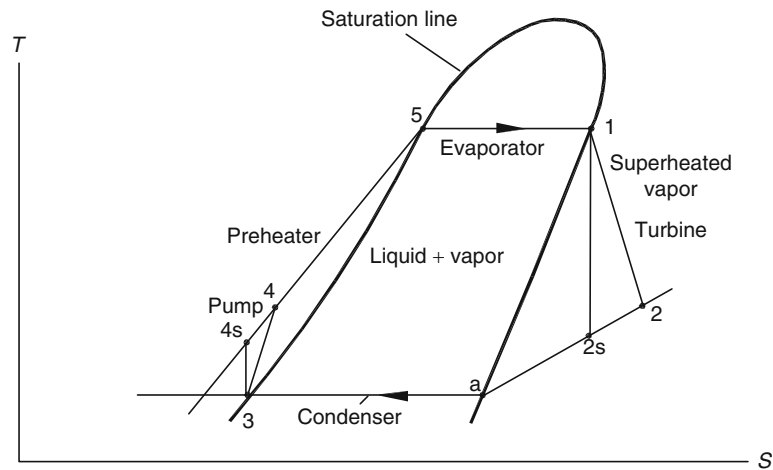
condenser and is pumped back to the preheater via a feed pump.

Turbine Analysis The vapor expands in the turbine between points 1 and 2 of Fig. 20.



Geothermal Power Conversion Technology. Figure 19

Simplified schematic of an air-cooled binary geothermal power station [36]



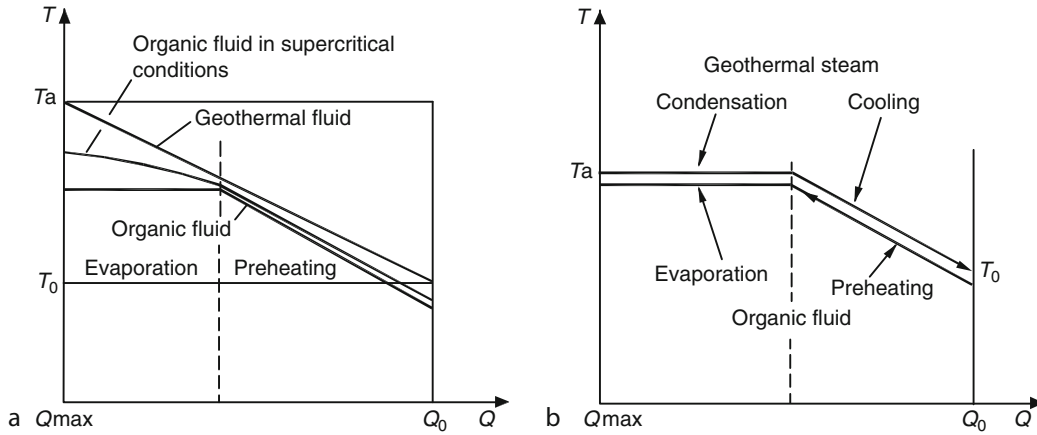
Geothermal Power Conversion Technology. Figure 20

T-S diagram showing a basic Organic Rankine cycle

Assuming the usual of steady adiabatic operation, the power is determined from:

$$\dot{W}_t = \dot{m}_{wf}(h_1 - h_2) = \dot{m}_{wf}\eta_t(h_1 - h_{2s}) \quad (40)$$

where η_t is the isentropic turbine efficiency (known parameter – the organic working fluid \dot{m}_{wf} expands into the superheated zone). For a given working fluid, the thermodynamic properties can be found from fluid



Geothermal Power Conversion Technology. Figure 21

T-Q diagrams of Organic Rankine cycle operated by (a) liquid and (b) geothermal steam

tables for selected design parameters. Selection of the turbine power output then helps determining the required working fluid mass flow rate.

Feed Pump Analysis Using similar assumptions as for the other components, the power imparted to the working fluid from the feed pump (points 3–4 of Fig. 20) is:

$$\dot{W}_p = \dot{m}_{wf}(h_4 - h_3) = \dot{m}_{wf}(h_{4s} - h_3)\eta_p \quad (41)$$

where η_p is the isentropic pump efficiency.

Condenser Analysis Condenser heat rejection occurs between points 2 and 3 on the cycle diagram in of Fig. 20).

The heat that must be rejected from the working fluid to the cooling medium, either water (shown here) or air, is given by:

$$Q_c = \dot{m}_{wf}(h_2 - h_3) \quad (42)$$

The relationship between the flow rates of the working fluid and the cooling water is:

$$\dot{m}_{cw}(h_{out} - h_{in}) = \dot{m}_{wf}(h_2 - h_3) \quad (43)$$

Assuming a constant specific heat \bar{c} for cooling water temperature between T_{out} and T_{in} , the cooling water mass flow rate can be found from:

$$\dot{m}_{cw}\bar{c}(T_{out} - T_{in}) = \dot{m}_{wf}(h_2 - h_3) \quad (44)$$

or

$$\dot{m}_{cw} = \dot{m}_{wf} \frac{(h_2 - h_3)}{\bar{c}(T_{out} - T_{in})} \quad (45)$$

Equation 44 will change to:

$$\dot{m}_{air}(h_{airout} - h_{airin}) = \dot{m}_{wf}(h_2 - h_3) \quad (46)$$

Assuming constant specific heat C_p for air in the relevant temperature range and neglecting humidity influence then:

$$\dot{m}_{air}C_{p,air}(T_{airout} - T_{airin}) = \dot{m}_{wf}(h_2 - h_3) \quad (47)$$

or

$$\dot{m}_{air} = \dot{m}_{wf} \frac{(h_2 - h_3)}{C_{p,air}(T_{airout} - T_{airin})} \quad (48)$$

Preheater and Evaporator Analysis Heat transfer to the working fluid takes place between points 4 and 1 in Fig. 20. Firstly, there is preheating between points 4 and 5, followed by evaporation up to point 1.

Secondly, on the brine side there is a continuous cooling due to the heat transfer to the organic fluid. Viewing the entire package as the thermodynamic system, the prevailing equation is:

$$\dot{m}_b(h_a - h_c) = \dot{m}_{wf}(h_1 - h_4) \quad (49)$$

If the brine is considered as liquid only, then the left-hand side of the equation can be replaced by the

average specific heat of the brine \bar{c}_b multiplied with the temperature drop:

$$\dot{m}_b \bar{c}_b (T_a - T_c) = \dot{m}_{wf} (h_1 - h_4) \quad (50)$$

The following equation can be used to find the required brine flow rate for a given set of cycle design parameters:

$$\dot{m}_b = \dot{m}_{wf} \frac{h_1 - h_4}{\bar{c}_b (T_a - T_c)} \quad (51)$$

The total energy transfer between the brine and the organic vapor takes place between clearly between points 4 and 1 shown in Fig. 22. This data is required for the design of the individual heat exchangers. The abscissa represents the total amount of heat passed from the brine to the working fluid and can be presented either in percent or in total heat flow per unit time (kJ/h).

The preheater provides sensible heat to raise the working fluid to its boiling point, state 5. The evaporation occurs from 5 to 1 along an isotherm for a pure working fluid. The point in the heat exchanger where the brine and the working fluid experience the minimum temperature difference is called the pinch-point

and is designated the pinch-point temperature difference ΔT_{pp} (see Fig. 22).

State points 4, 5, and 1 should be known from the cycle specifications. State 4 has the values of the compressed liquid at the outlet from the feed pump, state 5 is a saturated liquid at the boiler pressure, state 1 is a saturated vapor (same as at the turbine inlet condition). The geoliquid transfers heat to the evaporator from point a to point b, with the rest in the preheater down to point c. The two heat exchangers may be analyzed separately as follows:

$$\text{Preheater : } \dot{m}_b \bar{c}_b (T_b - T_c) = \dot{m}_{wf} (h_5 - h_4) \quad (52)$$

$$\text{Evaporator : } \dot{m}_b \bar{c}_b (T_a - T_b) = \dot{m}_{wf} (h_1 - h_5) \quad (53)$$

The brine inlet temperature T_a is always known. The pinch-point temperature difference is known from manufacturer's specifications. This allows T_b to be determined from the known value for T_5 .

The evaporator heat transfer surface is between the two fluids A_E , and can be determined from the basic heat transfer relationship:

$$Q_E = \bar{U} A_E LMTD|_E \quad (54)$$

where \bar{U} is the overall heat transfer coefficient and $LMTD|_E$ is the log mean temperature difference.

For detailed calculations see DiPippo [33].

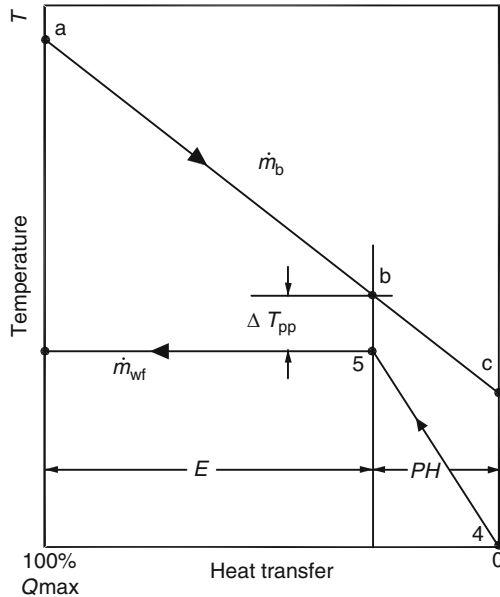
Since heat exchangers can be built in a variety of geometrical arrangements (shell-and-tube, plate, parallel flow, counter flow, etc.), there are correction factors to be used with the equations given above depending on the configuration. Refer to [37], for more details.

Overall Cycle Analysis Organic Rankine cycle performance can be assessed by the First Law using the thermal efficiency:

$$\eta_{th} \equiv \frac{\dot{W}_{net}}{\dot{Q}_{PH+E}} \quad (55)$$

Since the net power of the cycle is the difference between the thermal power input and the rejected thermal energy, this formula can be rewritten as:

$$\eta_{th} = \frac{\dot{Q}_{PH+E} - \dot{Q}_t}{\dot{Q}_{PH+E}} = 1 - \frac{\dot{Q}_c}{\dot{Q}_{PH+E}} = 1 - \frac{h_2 - h_3}{h_1 - h_4} \quad (56)$$



Geothermal Power Conversion Technology. Figure 22 Temperature-heat transfer diagram for preheater and evaporator

ORC efficiency is low because of low source temperatures and after subtracting all the parasitic loads from the gross power output, the final cycle efficiency may result in about 10%.

Another measure of station performance can be obtained using the Second Law in the form of the utilization efficiency η_u , which is defined (see Eq. 38) as the ratio of the actual net station power to the maximum theoretical power obtainable from the geothermal fluid in the reservoir state:

$$\eta_u = \frac{\dot{W}_{\text{net}}}{\dot{E}_{\text{res}}} = \frac{\dot{W}_{\text{res}}}{\dot{m}_b[(h_{\text{res}} - h_0) - T_0(s_{\text{res}} - s_0)]} \quad (57)$$

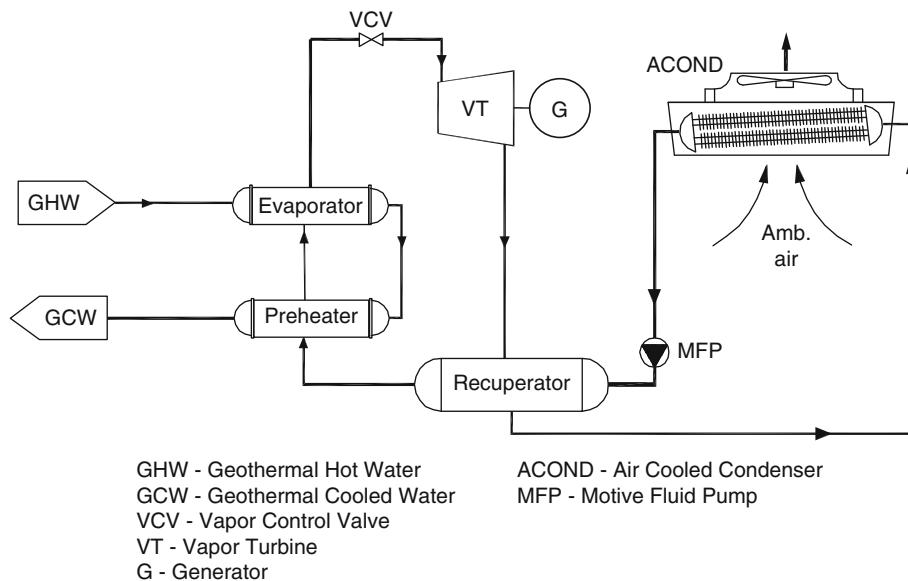
where \dot{E}_{res} is \dot{W}_{max} defined in the section on “Overall Thermal Efficiency,” T_0 is the dead-state temperature (the local wet-bulb temperature if a water cooling tower is used), h_0 and s_0 are the enthalpy and entropy values for the geothermal fluid evaluated at the dead-state pressure and temperature (usually approximated as the saturated liquid values at T_0).

Recuperated Organic Rankine Cycle The efficiency of the Organic Rankine cycle described in the section on “Organic Rankine Cycle-based Power Generation Process” can be improved by recovering part of the

heat of the superheated vapor before it enters the condenser (T_2 to T_a in Fig. 20) [53]. This is particularly important when there is limitation in the cooling temperature of the brine and condensate mixture. The silica scaling risk is the limiting factor in most of cases. It is increased as the brine temperature drops. In this case, the recuperator provides some of the preheating heat from the vapor, exiting the turbine.

The recuperator is applicable when the organic fluid is of the “dry expansion” type, where the expansion in the turbine is done in the dry superheated zone and the expanded vapor contains heat that has to be extracted prior to the condensing stage (Figs. 23 and 24). The recuperated Organic Rankine cycle is typically 10–15% more efficient than the simple Organic Rankine cycle described at the beginning of this chapter). This applies also to the two-phase geothermal power station as given in Figs. 25 and 26.

Temperature Cascading Organic Rankine Cycle A cascading system can be used to increase the power output of a binary power station [43]. In a simple cascading method there are two or more evaporators and preheaters, arranged in consecutive structure. The geothermal fluid travels from one pair of units to the

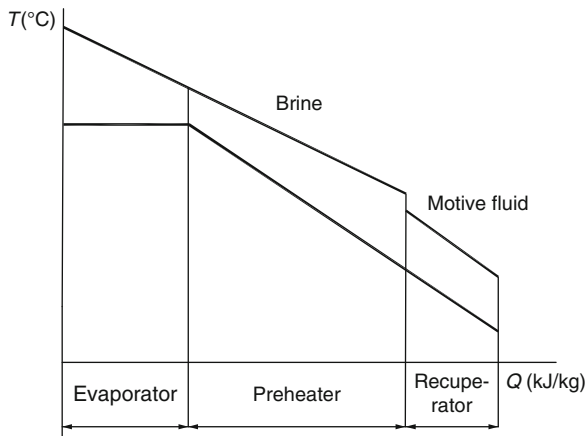


Geothermal Power Conversion Technology. Figure 23

Recuperated Organic Rankine cycle in simple binary power station – schematic

other. The station incorporates three sets of organic systems, each working in different ranges of temperatures. In an improved cascading design, the evaporators are arranged in series while the preheaters all work in the same temperature range [38]. Schematic design is given in Fig. 27.

The T-Q diagram describing this combination is in Fig. 28. Since evaporation is performed at three



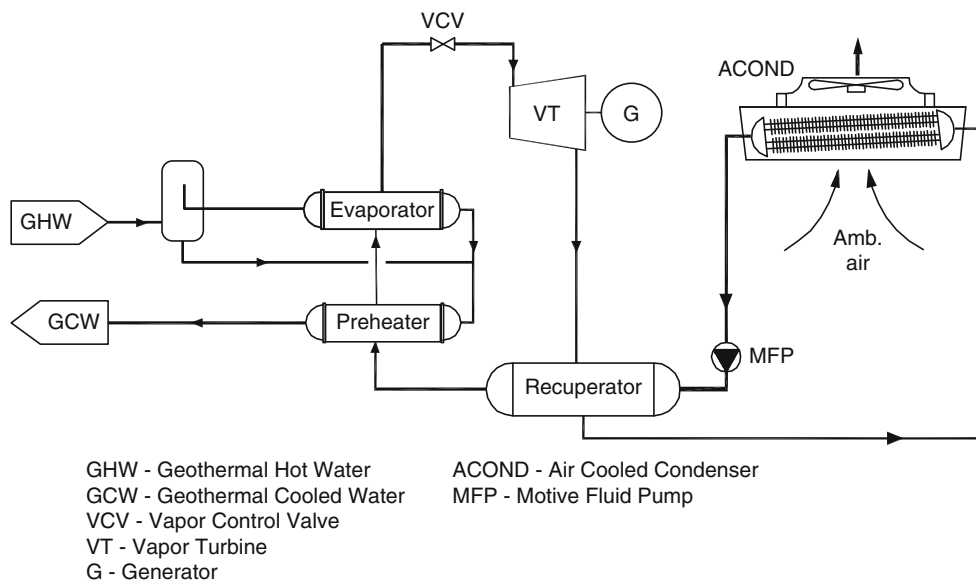
Geothermal Power Conversion Technology. Figure 24
Recuperated Organic Rankine cycle in simple binary power station

different temperatures, three turbines are required for such an operation.

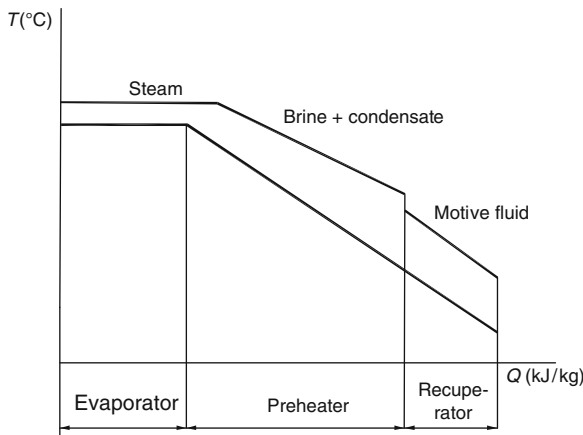
Dual-Pressure Organic Rankine Cycle A dual-pressure cycle is designed to reduce the thermodynamic losses incurred in the brine heat exchangers of the basic cycle. These losses are due to the transferring heat across a large temperature difference between the hotter brine and the cooler working fluid, see Fig. 29. By maintaining a closer match between the brine cooling curve and the working fluid heating/boiling curve, these losses can be reduced.

The dual-pressure cycle has a two-stage heating/boiling process that allows the two fluids to achieve a smaller average temperature difference than the one-stage process used in a basic cycle. A dual-pressure station schematic is given in Fig. 29 and the corresponding process diagram is shown in Fig. 30.

A dual-admission turbine is required to allow low-pressure saturated vapor (state 9) to be admitted to the turbine to mix with the partially expanded high-pressure vapor (state 2) to form a slightly superheated vapor (state 3). Given the small size of turbines using organic working fluids, practical considerations may lead to an alternative design using two separate turbines.



Geothermal Power Conversion Technology. Figure 25
Recuperated ORC in two-phase binary power station – schematic



Geothermal Power Conversion Technology. Figure 26
Recuperated ORC in two-phase binary power station

The analysis of a dual-pressure cycle follows the same methodology as a basic cycle but takes more time. A detailed comparison of basic cycles (single-pressure) and the dual-pressure cycles was conducted by Khalifa and Rhodes [39] for two different working fluids. Their results show that in all cases, the thermal efficiency for a dual-pressure cycle is lower than for a basic cycle. However, the utilization efficiency for a dual-pressure cycle is significantly higher than for a basic cycle, ranging from a 6% advantage at the highest brine temperature to 24% advantage at the lowest.

The explanation for this is that thermal efficiency depends on the amount of heat added to the cycle but makes no distinction between resources maximum energy, ignoring the temperature difference between the fluids.

However, the utilization efficiency depends on how effectively the energy of the brine is used. By more closely matching the brine cooling curve to the heating and boiling curves, the average temperature difference between the two fluids is decreased and the irreversibilities are reduced. This allows more energy from the brine to enter the cycle leading to a higher overall utilization efficiency.

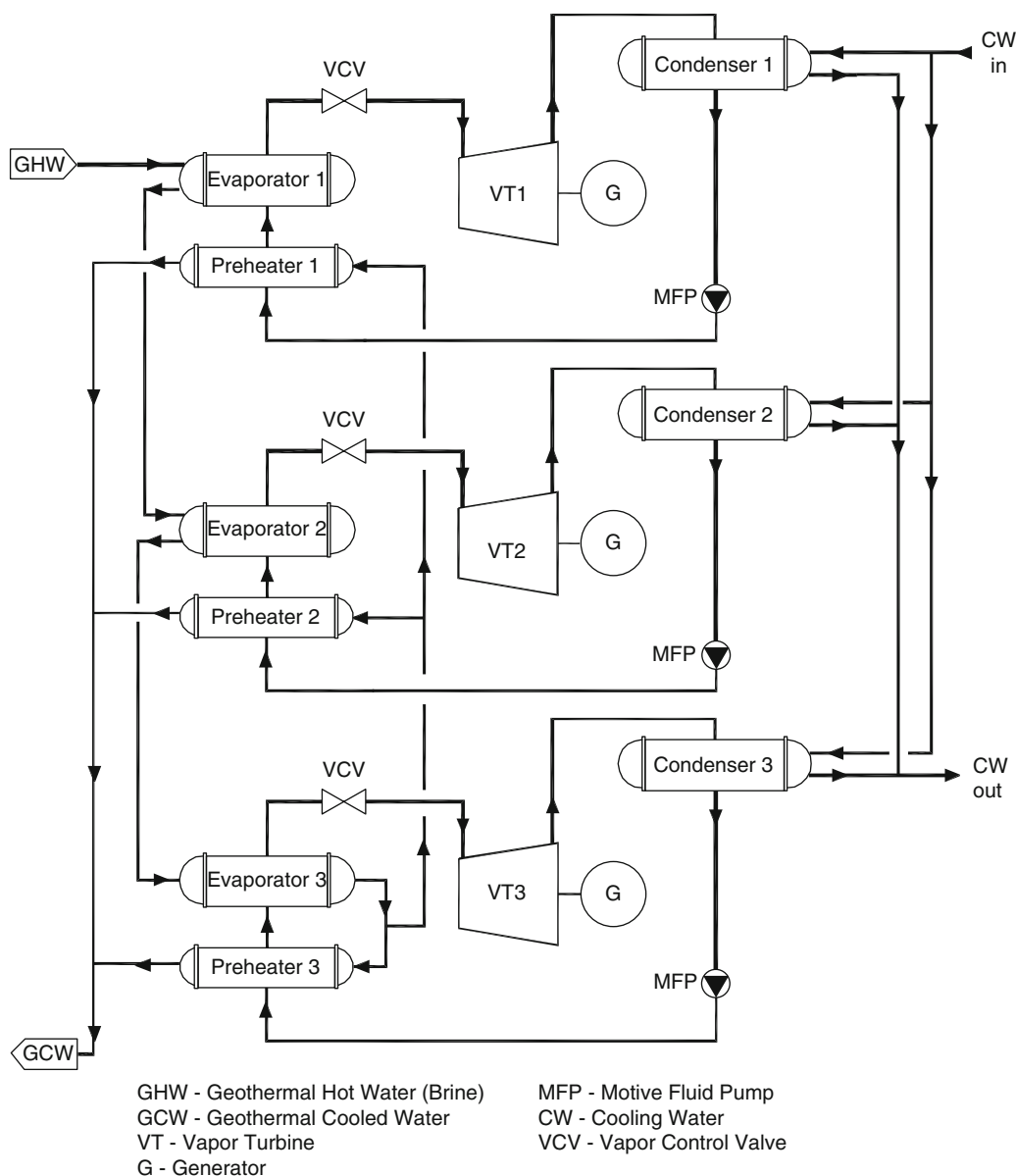
The 5 MW Raft River Dual-Boiling station in Idaho, USA, was the first to make use of the dual-pressure concept [40] and was operated as a demonstration station from 1981 to 1982 by the Idaho National Engineering Laboratory for the US Department of Energy.

Supercritical Organic Rankine Cycles As mentioned in section on “[Geothermal Resources](#)” and shown in [Fig. 21a](#), an Organic Rankine Cycle where the operating liquid in the supercritical zone follows the heat source cooling curve more closely and the irreversibility losses of heat transfer are reduced as discussed by Tester and Milora [41] in detail. Four different cycles were examined. The operating fluid selected to illustrate the process is R-115 (C_2ClF_5) having a critical temperature of $80^\circ C$ and being a suitable working fluid for both subcritical and supercritical operation at a geothermal fluid temperature of $150^\circ C$. In all four cases, the vapor was heated to $135^\circ C$ and condensed at $26.7^\circ C$.

The first cycle is performed at 27.5 bar, i.e., subcritical cycle. The cycle efficiency is 9%, utilization efficiency is 46.5% and feed pump pressure ratio is 0.87. There is a quite large temperature gap between the cooling line of the geothermal fluid and the heated fluid. To improve the heat transfer, the second cycle pressure was increased to 39.26 bar which is already in the supercritical zone. This increased the cycle efficiency to 11.2% and utilization efficiency to 56.5% but feed pump pressure ratio increased to 1.24. Further increase of the pressure to 80 bar resulted in almost parallel cooling and heating lines. This increased both efficiencies to 11.9% and 63.2% respectively, and feed pump pressure ratio to 2.54. However, additional increase to 114.4 bar dropped cycle efficiency to 10.6% and utilization efficiency to 54.6%, feed pump pressure ratio to 3.62 and accompanied by expansion through the wet zone meaning a possible droplets impingement on the turbine blades.

In summary, the work at supercritical condition has improved the cycle and utilization efficiencies but simultaneously substantially increasing feed-pump work. There is a large pumping power requirement to nearly 50% of the net turbine power. Although the operation in 80 bar seems to achieve the highest cycle efficiency and highest utilization, the question of cost of an 80 bar structure and operation in high pressure against efficiency benefits will influence the final selection of operating conditions.

Nevertheless in more moderate conditions, supercritical cycles have been used, but the cycle efficiency improvement is impaired by the increased cycle pump losses. See the section on “[Efficiency and Work Ratio](#).”

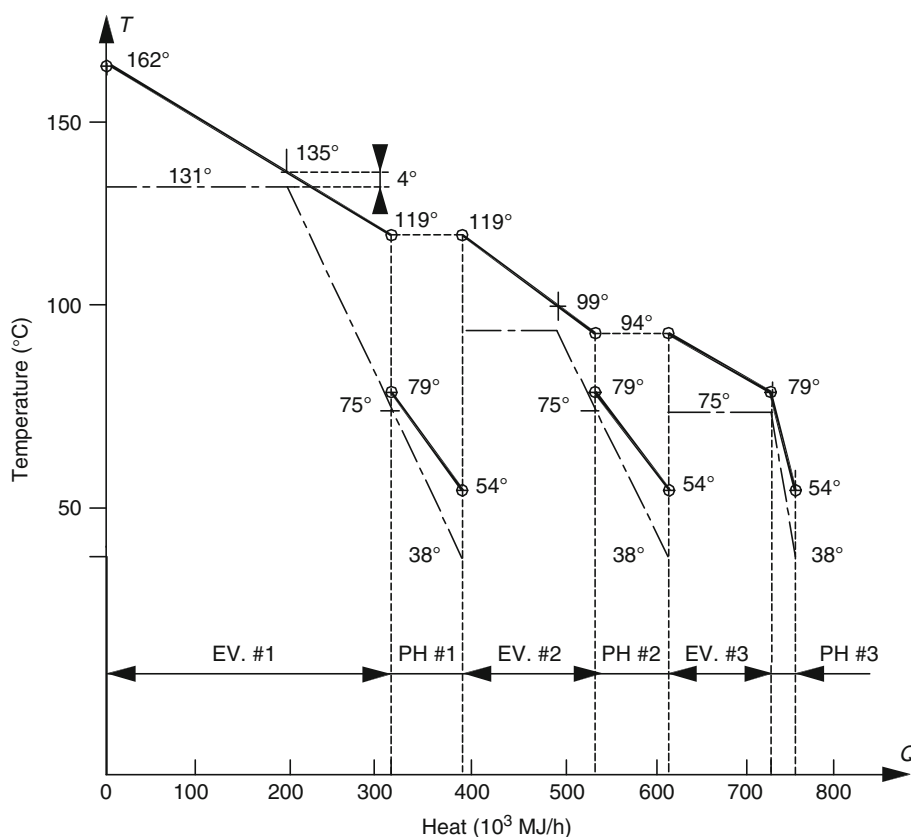


Geothermal Power Conversion Technology. Figure 27
Cascading ORC schematics

Dual-Fluid Organic Rankine Cycle Experimental – see section on “[Experimental Power Stations](#).”

Working Fluid Selection Selection of the appropriate working fluid has great bearing on the performance of the Organic Rankine cycle. Considerations should include both thermodynamic properties of the fluids as well as health, safety, and environmental impact [51, 52].

Thermodynamic Properties [Table 3](#) lists some candidate fluids and their relevant thermodynamic properties. Pure water is included for comparison [42]. Clearly all of the candidate fluids have critical temperatures and pressures far below water. Furthermore, since the critical pressures are reasonably low, it is feasible to consider supercritical cycles for the hydrocarbons. As already mentioned in “[Supercritical Organic Rankine Cycles](#),” this allows a better match



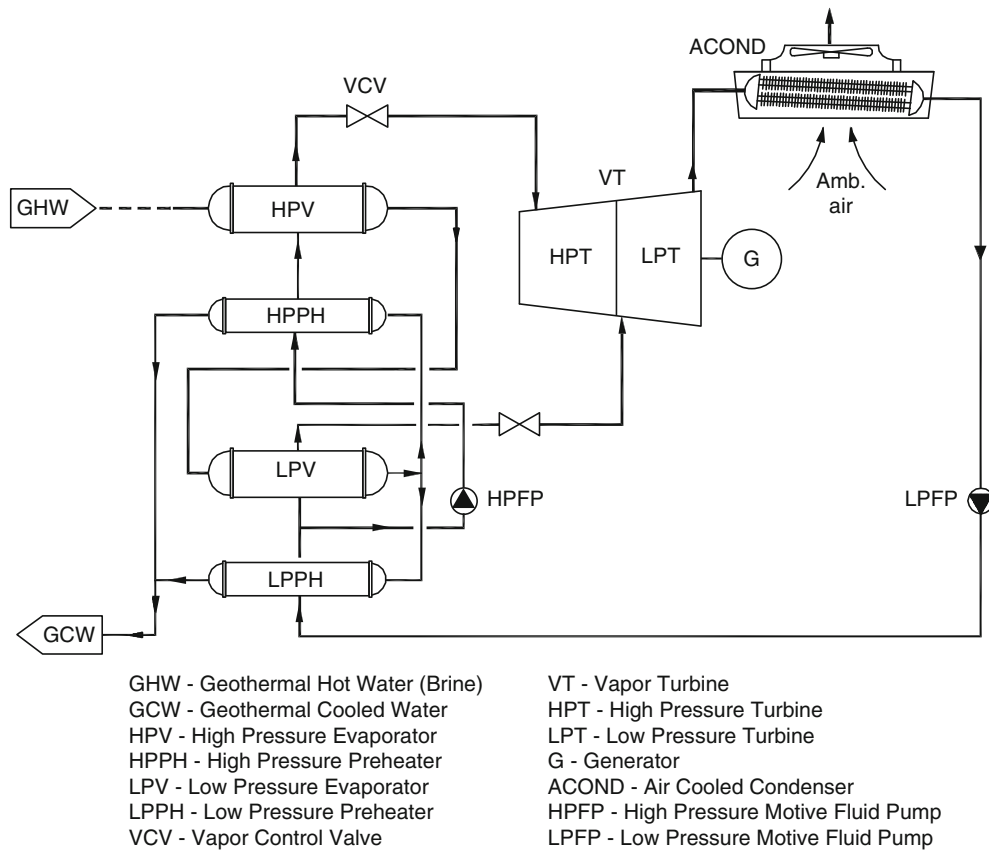
Geothermal Power Conversion Technology. Figure 28
T-Q diagram for cascading ORC

between the brine cooling curve and the working fluid heating-boiling line, reducing the thermodynamic losses in the heat exchangers. However, the net power output may not be higher resulting from the high pumping energy required by the supercritical cycle.

Mixtures of these fluids have been studied for use in geothermal binary stations. In particular, the thermodynamic properties of 90% C_4H_{10} and 10% $\text{i-C}_5\text{H}_{12}$ were determined by the National Bureau of Standards (predecessor of NIST) in Washington [43] when it was chosen as the working fluid for the Heber Binary Demonstration station in the 1980s [44]. Mixtures evaporate and condense at variable temperature, unlike pure fluids that change phase at constant temperature. This means that subcritical pressure boilers for mixed fluids can be better matched to the brine curves, similar to, but not exactly like supercritical pure fluids. A practical hurdle in the use of water ammonia (Kalina cycle) mixture is that the differential leaks of the two fluids

in various points of the system modify mixture composition over time.

Another important characteristic of candidate fluids is the shape of the saturated vapor curve as viewed in temperature–entropy coordinates, see Fig. 31. This curve for water (thin line) has a negative slope, but certain hydrocarbons and refrigerants show a positive slope for portions of the saturation line. That is, a local minimum in the entropy at some low temperature exists, T_m and local maximum in entropy at higher temperature, T_M . Retrograde fluids include normal and iso-butane, normal and iso-pentane. These fluids exhibit retrograde behavior over the following temperature ranges, $T_m \rightarrow T_M$: C_4H_{10} , $-3^\circ\text{C} \rightarrow 127^\circ\text{C}$; $\text{i-C}_4\text{H}_{10}$, $-3^\circ\text{C} \rightarrow 117^\circ\text{C}$; C_5H_{12} , $-3^\circ\text{C} \rightarrow 177^\circ\text{C}$; and $\text{i-C}_5\text{H}_{12}$, $-13^\circ\text{C} \rightarrow 177^\circ\text{C}$. Since T_m is lower than any temperatures encountered in geothermal binary stations, these fluids can be taken as having saturated vapor lines similar to that shown in Fig. 32 for practical



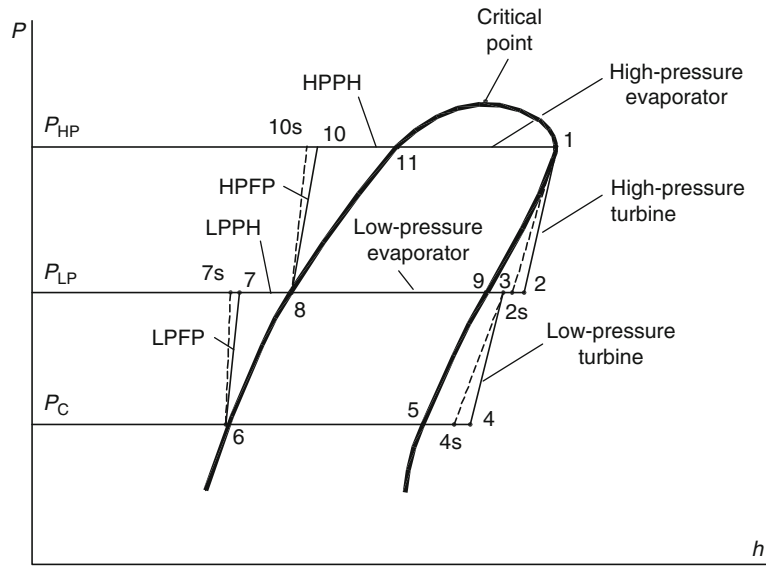
Geothermal Power Conversion Technology. Figure 29

Dual-pressure binary station based on single turbine with LP and HP inlets. Simplified flow diagram

purposes. This has major implications for Rankine cycles.

On the one hand, normal fluids such as water require considerable superheating, extending the isobar a–b–c upward to avoid excessive moisture at the turbine exhaust, state g. On the other hand, retrograde (backbend) fluids allow expansion from the saturated vapor line into the superheated region, process b–f, avoiding any moisture during the turbine expansion process. It has been shown [45] that it is possible to run a supercritical cycle where the turbine inlet state lies above the critical point and the expansion line lies inside the wet region for a portion of the process, emerging into the superheated region, without suffering any wetness penalty in efficiency. Apparently, the fluid remains in a metastable vapor state while passing through the wet region by staying on the dry side of the Wilson line [46].

Turbine Size While evaluating the potential working fluid, turbine size and cost must also be considered as part of that task. Milora and Tester [47] compared a line of hydrocarbons that are suitable for use in binary systems (including steam for comparison) on the basis of nondimensional turbine design parameters. They used four basic parameters: turbine blade diameter, turbine rotational speed, enthalpy drop, and volumetric flow rate. Using the factors that build those parameters, they established a “figure of merit” that is influenced by the fluid molecular weight, heat of evaporation (h_{fg}), specific volume, critical pressure, etc. Additional evaluation was made on the exit flow area that is also influenced by the same factors as shown by them [47] and also by DiPippo [33]. The results of both evaluations show that for the same power output and temperature range, systems that use low molecular weight fluids like ammonia (NH_3) result in a smaller turbine than butane or pentane by



Geothermal Power Conversion Technology. Figure 30
Pressure–Enthalpy (P-h) diagram of a dual-pressure binary station

Geothermal Power Conversion Technology. Table 3 Thermodynamic properties of some candidate working fluids for binary stations

Fluid	Formula	T_c (°C)	P_c (MPa)	P_c (Lbf/in ²)	P_s @ 300 K (MPa)	P_s @ 400 K (MPa)
Propane	C ₃ H ₈	96.95	4.236	614.4	0.9935	n.a.
i-Butane	i-C ₄ H ₁₀	135.92	3.685	534.4	0.3727	3.204
n-Butane	C ₄ H ₁₀	150.8	3.718	539.2	0.2559	2.488
i-Pentane	i-C ₅ H ₁₂	187.8	3.409	494.4	0.09759	1.238
n-Pentane	C ₅ H ₁₂	193.9	3.240	469.9	0.07376	1.036
Ammonia	NH ₃	133.65	11.627	1,686.3	1.061	10.3
Water	H ₂ O	374.14	705.45	3,203.6	0.003536	0.24559

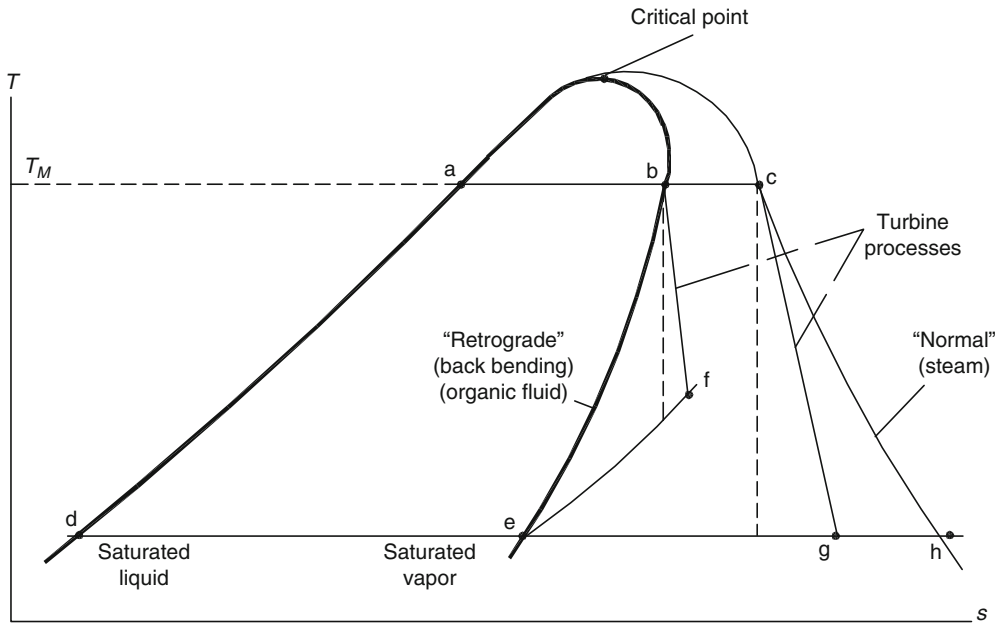
factors of 5 and 12 respectively, and for comparison smaller than steam turbine by factor of 120–150 times due to the large specific volume of steam in low temperatures. These general evaluations do not eliminate the need for particular turbine design for actual operating conditions.

Environmental Safety and Health Considerations The fluids used in the Organic Rankine cycle should comply with local and international health and environmental rules and agreements such as the Copenhagen Amendment of 1994 and Montreal Protocol of 1987 (in

comparison with R-12 and R-114 that were banned from use and rated 1 for Ozone Depletion Potential (ODP)). Other potential fluids in Table 4 [48] are acceptable even though they are flammable. The table also compares Global Warming Potential which illustrates main reason for banning R-12 and R-114.

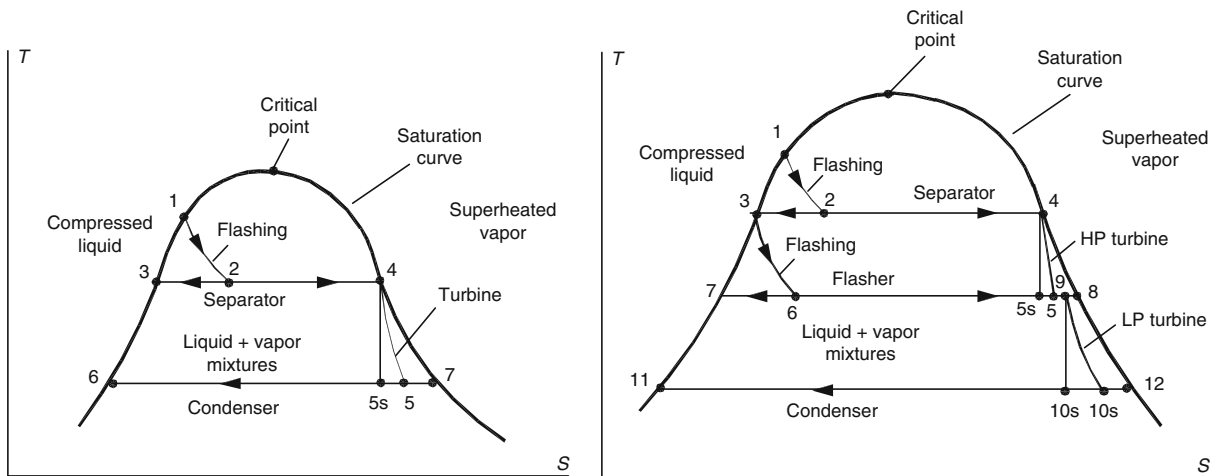
Ancillary Systems

Separation Process The separation process is considered as one at constant pressure, i.e., an isobaric process once the flash has taken place. The quality or



Geothermal Power Conversion Technology. Figure 31

Temperature-entropy diagram contrasting normal and retrograde saturated vapor curves



Geothermal Power Conversion Technology. Figure 32

T-S diagrams for single (left) and double-flash (right) steam cycles

dryness fraction, x , of the mixture that forms after the flash, state 2 (Fig. 32), can be found from:

$$x_2 = \frac{h_2 - h_3}{h_4 - h_3} = \frac{h_1 - h_3}{h_4 - h_3} = \frac{h_1 - h_3}{h_{fg}(T_3, P_3)} \quad (58)$$

(h_{fg} is latent heat of evaporation (B)) by using the “lever” rule between points 3 and 4, the steam mass fraction of the mixture can be found. It specifies the amount of steam flowing to the turbine per unit total mass of flow into the separator.

For the expansion, condensing and cycle analysis refer to the section on “Expansion Process.”

Geothermal Power Conversion Technology. Table 4 Environmental and health properties of some candidate working fluids [41]

Fluid	Formula	Toxicity	Flammability	ODP	GWP
R-11	CCl_3F	Nontoxic	Non-flam	1.0	4,000
R-12	CCl_2F_2	Nontoxic	Non-flam	1.0	4,500
R-113	CCl_3CF_3	Nontoxic	Non-flam	0.8	4,800
R-114	$\text{C}_2\text{Cl}_2\text{F}_4$	Nontoxic	Non-flam	0.7	5,850
Propane	C_3H_8	Low	Very high	0	3
i-Butane	$\text{i-C}_4\text{H}_{10}$	Low	Very high	0	3
n-Butane	C_4H_{10}	Low	Very high	0	3
i-Pentane	$\text{i-C}_5\text{H}_{12}$	Low	Very high	0	3
n-Pentane	C_5H_{12}	Low	Very high	0	3
Ammonia	NH_3	Toxic	Lower	0	0
Water	H_2O	Nontoxic	Non-flam	0	–
R245fa	$\text{CF}_3\text{CH}_2\text{CHF}_2$	Nontoxic	Non-flam	0	950

Flashing Process The sequence of processes begins with geothermal fluid under pressure at state 1, close to the saturation curve. The flashing process is modeled as one at constant enthalpy, or isenthalpic process as it occurs steadily, spontaneously, essentially adiabatically, and with no work involvement also ignoring any change in the kinetic or potential energy of the fluid as it undergoes the flash, therefore:

$$h = \text{constant} \quad (59)$$

Flash and Separation Processes Referring to Fig. 32b, the two flash processes 1–2 and 3–6 are analyzed the same as the flash process for the single-flash station in Figs. 13 and 32a. Each process generates a fractional amount of steam given by the quality x , of the 2-phase mixture. Each flash is followed by a separation process. The governing equations are as follows:

$$h_1 = h_2 \quad (60)$$

$$x_2 = \frac{h_2 - h_3}{h_4 - h_3} \quad (61)$$

$$h_3 = h_6 \quad (62)$$

$$x_6 = \frac{h_3 - h_7}{h_8 - h_7} \quad (63)$$

The mass flow rates of the steam (\dot{m}_{hps}) and liquid (brine) (\dot{m}_{hpb}) for the high- and low-pressure stages are found from:

$$\dot{m}_{\text{hps}} = x_2 \dot{m}_{\text{total}} = \dot{m}_4 = \dot{m}_5 \quad (64)$$

$$\dot{m}_{\text{hpb}} = (1 - x_2) \dot{m}_{\text{total}} = \dot{m}_3 = \dot{m}_6 \quad (65)$$

$$\dot{m}_{\text{lps}} = (1 - x_2) x_6 \dot{m}_{\text{total}} = \dot{m}_8 \quad (66)$$

$$\dot{m}_{\text{lpb}} = (1 - x_2)(1 - x_6) \dot{m}_{\text{total}} = \dot{m}_7 \quad (67)$$

These mass flows will be used to calculate the power generated from the two stages of turbine expansion, the amount of waste liquid to be disposed of and the heat that must be rejected through the condenser and ultimately from the cooling tower.

Optimization

Optimum Wellhead Pressure Once a valve is installed on the well, the pressure at which the power station is to operate must be determined with the wellhead pressure being controlled by a throttling valve. The well productivity curve can be approximated as an elliptical equation in terms of the mass flow rate of steam as a function of the wellhead pressure:

$$\left[\frac{\dot{m}}{\dot{m}_{\text{max}}} \right]^2 + \left[\frac{P}{P_{\text{ci}}} \right]^2 = 1 \quad (68)$$

where \dot{m}_{\max} is the maximum observed mass flow rate at full open valve and P_{ci} is the closed-in wellhead pressure. This function is shown schematically in Fig. 33. Assuming that values for these two parameters are available from well tests, the mass flow rate at any pressure can be calculated by:

$$\dot{m} = \dot{m}_{\max} \sqrt{1 - (P/P_{ci})^2} \quad (69)$$

Since opening the wellhead valve is a throttling process, the enthalpy of the steam remains the same.

Turbine power is proportional to the product of the steam mass flow rate and the enthalpy drop Δh (shown as an ideal isentropic process), from $h_{ci} = h_1$, down to assumed condenser pressure. The maximum is located somewhere in between.

Compute and solve for the power output per maximum steam flow rate by using Eqs. 15 and 69 as follows:

$$\frac{\dot{W}}{\dot{m}_{\max}} = \frac{\dot{W}}{\dot{m}} \times \frac{\dot{m}}{\dot{m}_{\max}} = (h_1 - h_2) \times \sqrt{1 - (P/P_{ci})^2} \quad (70)$$

where $(h_1 - h_2)$ is the isentropic enthalpy drop across the turbine (Δh in Fig. 34), that can be obtained graphically using a large-scale Mollier diagram, see *Steam Tables* [32].

The work method involves preparing a table of Eq. 70 variables, i.e., h_1 , h_2 (for given condenser

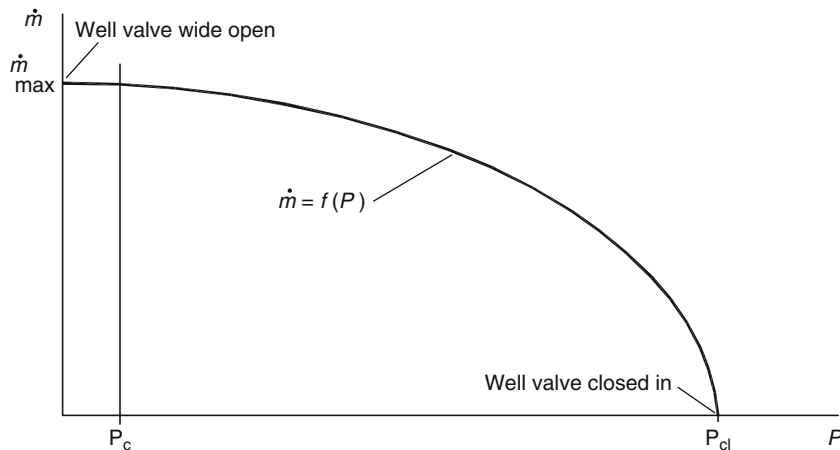
pressure), $h_1 - h_2$, P/P_{ci} and $\frac{\dot{W}}{\dot{m}_{\max}}$ as function of the pressure p changing between closed and open valve position.

DiPippo [33] solved this problem for a non-isentropic turbine for given wellhead data. The result has a parabolic shape with maximum at about 40% of the closed valve pressure.

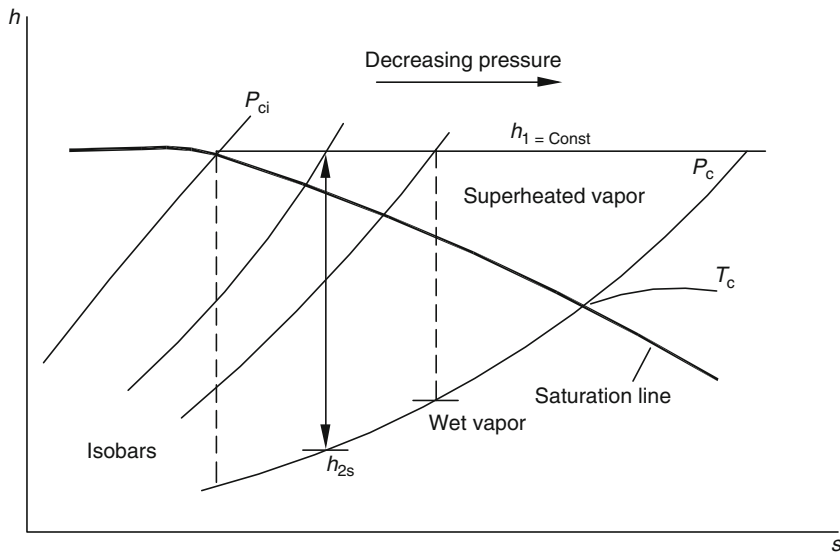
It should be noted that the curve is relatively flat near the optimum point, i.e., the power output is within 0.2% of the optimum, over a wide range of wellhead pressures. This allows for a wide enough pressure/valve setup range without sacrificing much of the power generation.

The optimization process for a double-flash station is more complicated than for a single-flash station due to the extra degree of freedom in the choice of operating parameters. This results in two maxima, one of which yields the highest power output and the other that has the best specific power output. The results are not identical.

An example of a double-flash optimization was made by DiPippo [49] for the Electric Power Research Institute. The specific and total power outputs reach their respective maxima at different points. The variation results in different second flashing temperatures and a difference of 2.5% in the thermal utilization efficiency. However, in comparison with a single-flash station operating in the same conditions the double-flash generates about 31% more power.



Geothermal Power Conversion Technology. Figure 33
Productivity curve for dry-steam wellhead



Geothermal Power Conversion Technology. Figure 34
Expansion from variable wellhead pressures (constant enthalpy)

If the rule of “equal temperature split” is used for initial evaluation, then the results for first and second flashing are close to the optimized calculations for maximum power and less accurate for maximum efficiency. This is acceptable for an approximation in view of the optimistic assumptions for the calculation that neglect pressure and thermal losses between the wellhead and the turbine.

Efficiency and Work Ratio The usual definition of thermal efficiency as the ratio between the net work done by the fluid and the total heat input to the cycle can be misleading in assessing the suitability of a given cycle in a heat engine. A concept of paramount importance in evaluating the suitability of a particular cycle for use in a heat engine is that of *work ratio*. This is defined as the ratio of the net work output of the cycle to the total positive (expansion) work of the cycle.

If there is very little negative work, as in a typical subcritical vapor cycle where only liquid of small specific volume has to be pumped at moderate pressure back into the boiler, the work ratio will be high. By contrast, this ratio is lower in a supercritical cycle where because of the high pressure, a larger portion of the positive work of the turbine is used to drive the feed pump [41].

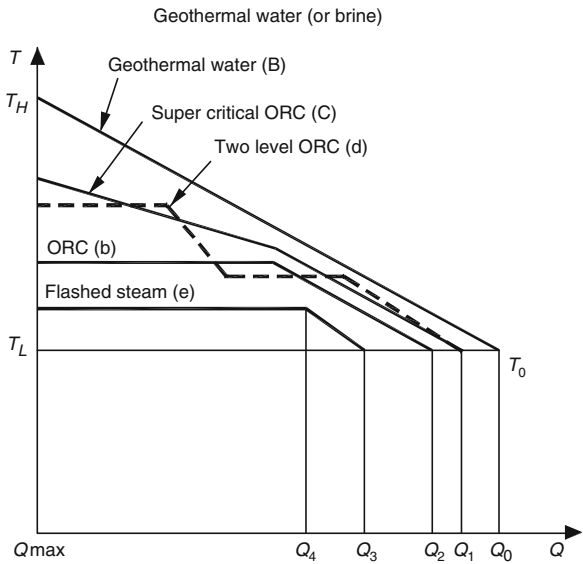
Taking all these practical implications of work ratio, it can be seen that the concept of work ratio into account, it can be seen that the concept of work ratio can be regarded as almost as important as the concept of ideal cycle efficiency in many ways. Refer to Table 5.

Optimizing the Efficiency by Matching the Cycle to the Heat Resource

Single-Phase Water or Brine Resource In Fig. 35, line (a) is a typical T-Q diagram showing the heat exchange between single phase water or brine geothermal fluid and Organic Rankine cycle fluid from the hydrocarbons family of materials. The geothermal fluid cools during the heat transfer operation while there are a few options for heat utilization by the Organic Rankine cycle fluid. The simple case is line (b) that comprises preheating and evaporation of the organic fluid. The optimal case would be a constant temperature difference between the two fluids. This is difficult to achieve, but is almost attained with Organic Rankine cycle fluid operation in a supercritical condition as show in line (c). However, the gain carries penalties. The first penalty is the large pumping power and the second is the high cost of hardware operating at such high pressures. Another option is using the Organic Rankine cycle fluid under the saturation curve applying step heating

Geothermal Power Conversion Technology. Table 5
Comparison of work ratio in supercritical and subcritical ORC

	Supercritical	Cascaded
	ORC	ORC
Gross (kW)	15,900	14,800
Fans (kW)	1,000	1,000
Feed Pump (kW)	2,450	900
Net	12,450	12,900
Work Ratio	78%	87%
For identical heat source and heat sink conditions:		
Heat source (liquid):		
Inlet temperature 170°C		
Outlet temperature 85°C		
Heat sink (air): 25°C		



Geothermal Power Conversion Technology. Figure 35
T-Q Diagram comparing of Organic and Flash Steam cycle for 160°C heat source

with one or more evaporation stages – line (d) for two stages or as in a cascading system described in Fig. 35 which gives a better fit.

Figure 35 line (e) illustrates the differences in the temperature drop between a Flash Steam Rankine

Geothermal Power Conversion Technology. Table 6
Comparison of power recovered by Ormat and Steam cycles for a 160°C heat source

Cycle comparison based on gross power		
	Steam	Organic
% preheat	9.6%	38%
Optimum Exit Temp	92.5	78
Heat Input	83%	100%
Thermal Efficiency (Gross)	10.7%	11.1%
Power (Gross)	80%	100%
Cycle comparison based on net power		
	Steam	Organic
% preheat	9.6%	38%
Optimum Exit Temp	92.5	78
Heat Input	83%	100%
Thermal Efficiency (Net)	10.6%	10.6%
Power (Net)	85%	100%

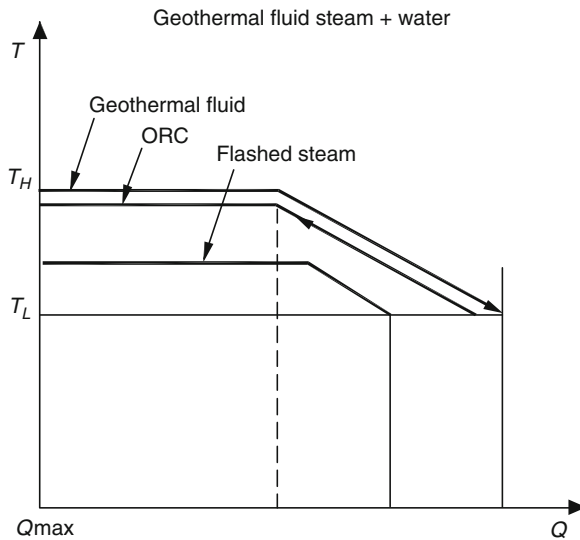
cycle and an Organic Rankine cycle. Because of the lower heat capacity of organic liquids and their much smaller latent heat of vaporization, these fluids lead to significantly smaller irreversibility losses of availability in the utilization of low or medium temperature predominantly sensible heat streams (Table 6).

Water-Dominated Two-Phase Flow When a substantial portion of the heat content of the geofluid is sensible heat of the liquid phase, the ORC has the advantage that its latent heat of evaporation is smaller than that of steam and therefore the vaporization temperature can be higher than that of a flash plant (Fig. 36). This leads to a higher efficiency and in addition eliminating the parasitic consumption of the condenser vacuum pumps.

Choked Well Flow The selection of separator operating conditions for liquid-dominated well has economical value. Optimization should be made for such evaluation. Two cases for choked and non-choked flows for liquid dominated wells are thoroughly studied by DiPippo [33, p. 98]. In his evaluation, two assumptions are made:

- There are no pressure losses between the wellhead and the turbine.
- Condensation occurs at certain known temperature (site specific).

As detailed in DePippo [33], the flow increases rapidly as the well is opened and the pressure is



Geothermal Power Conversion Technology. Figure 36
T-Q diagram of binary cycle driven by geothermal steam

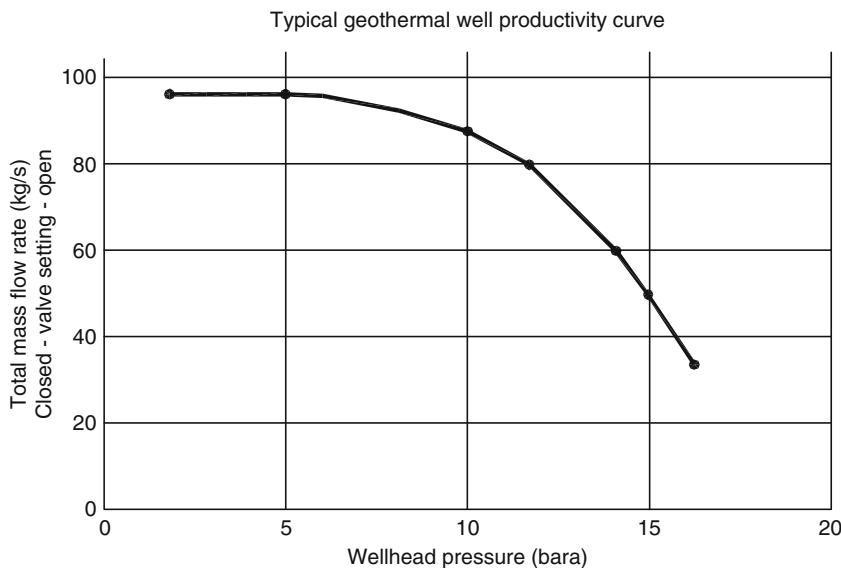
lowered. However, once the valve position is about 90% open (Fig. 37), the flow rate stabilizes and further opening of the well valve does not raise the total mass flow rate as the flow is “choked.”

The question now is what wellhead pressure should be chosen to maximize the power output from a single-flash station connected to this well.

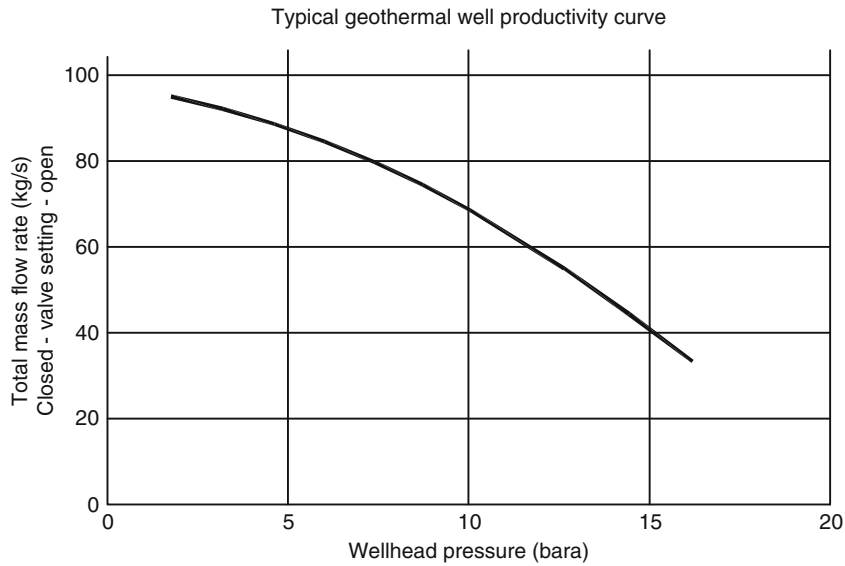
Eqs. 60–67 and 15–30 and a Mollier chart [32] are used to analyze the flashing, separation and turbine expansion processes. The calculations proceed in two phases:

- Phase 1 to determine the specific power output (measured by enthalpy drop on the Mollier chart) for a range of separator pressures (and equivalently temperatures).
- Phase 2 to find the total power by use of the variations of the total-flow rate as a function of the separator pressure.

Phase 1 calculations may rely on normal Steam Tables [31, 32] since the characteristics of the actual stream are unknown. Using the temperature and pressure of the maximum point, the actual flow rate is found from the productivity curve at that wellhead pressure. When this is multiplied by the corresponding specific power (Eq. 17) the maximum power can be obtained.



Geothermal Power Conversion Technology. Figure 37
Typical geothermal well productivity curve



Geothermal Power Conversion Technology. Figure 38
Typical geothermal well productivity curve

Non-Choked Well Flow Many wells do not have a near flat response nearing open valve position. This may be a result of the well bore diameter being too small. The well is characterized by the high slope production data curve as in Fig. 38 meaning a continuously increased flow rate with further opening of the valve on the wellhead.

The results of phase 1 calculations are the same as for the first case. The results of phase 2 calculations are such that the maximum specific power and maximum total power are not at the same temperature. The designer must select the best choice based on economical factors in a particular set of the wells.

An Approximate Formulation for Separation Temperature Along with the previous method described above, DiPippo [33] developed an approximate method that leads to an easy solution for T_3 which is:

$$T_{3,\text{opt}} = \frac{T_1 + T_6}{2} \quad (71)$$

(T_1 is the inlet temperature and T_6 is the condenser temperature)

Since this rule indicates that the temperature range between the reservoir and the condenser is divided into

two equal segments, this rule is sometimes called the “equal-temperature-split” rule. This approximate rule applies to all flash stations regardless of the number of flashing steps [50]. The rule is that the temperature difference between the reservoir and the first flash is equal to the temperature difference between the first flash and the second flash, as well as between the second flash and the condenser.

Main Power Station Components

Steam Turbines

General Turbines used in geothermal applications are made of corrosion-resistant materials due to the presence of gases such as hydrogen sulfide which induce stress corrosion and erosion due to droplets and entrained solids.

A dry-steam power station utilizes similar turbines as those used in a fossil fuel power station (Fig. 39).

The turbines are single-pressure units with impulse-reaction blades, either single-flow for smaller units or double-flow for large units above 50 MW as in Fig. 40. The condensers can be either direct-contact (barometric or low-level) or surface-type (shell-and-tube).

Direct Exhaust Steam Turbine Large turbines usually sit over the direct contact condenser maintaining minimum pressure losses at the turbine exit. For small units it is often advantageous to arrange the turbine and condenser side by side, for maintenance reasons.

Turbines for Dry Steam or Single Flash A typical power station flow diagram is detailed in Fig. 41. At each production well (PW), there is equipment to control and monitor geothermal fluid flow from the well to the station. This equipment includes:

- Wellhead valve – WV
- Silencer or rock muffler – S or RM



Geothermal Power Conversion Technology. Figure 39 Dry Steam 35 MW Franco Tossi turbine in Costa Rica (Courtesy of ICE, Costa Rica)

- Particle remover/purifier – PU
- Emergency relief valve – ERV
- Piping and instrumentation (pressure and temperature gauges)

If wellhead separators are used, the cyclone separator, CS is located close to the wellhead on the same pad. Also note that the NCG disposal used in the example below is of the steam ejector type.

The single-flash steam power station is the most common geothermal power industry system. DiPippo states [33] that as of May 2007, there were 159 units of this kind in operation in 18 countries around the world with single-flash stations accounting for about 32% of all geothermal stations, constituting over 42% of the total installed geothermal power capacity in the world. The unit power capacity ranges from 3 to 90 MW, and the average power rating is 25.3 MW per unit.

Turbines for single-flash units are typically rated at 25–50 MW and consist of 4–5 stages of impulse-reaction blades. Both single and double-flow designs are in use. Overall isentropic efficiencies in the high 80% range have been obtained.

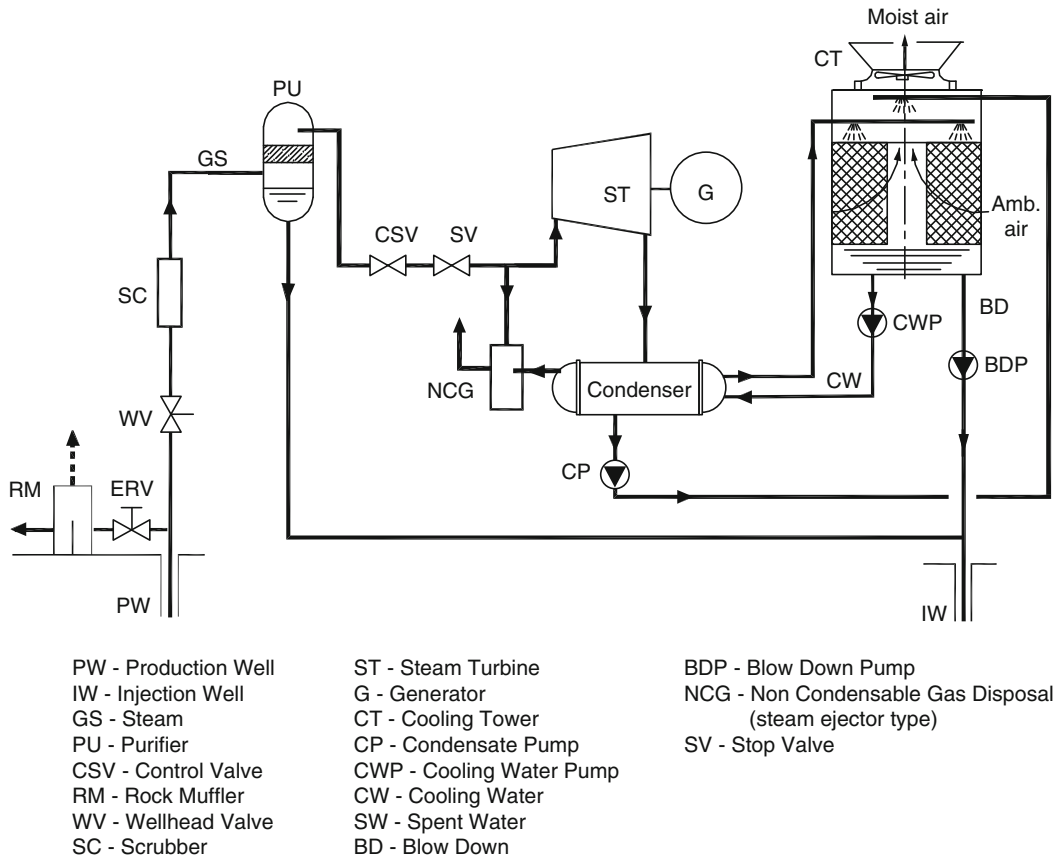
For the thermodynamic analysis of the energy conversion process, it is assumed that the geothermal fluid starts as a compressed liquid somewhere in the reservoir. It experiences a flashing process separating the two phases in the cyclone separator. The steam is used to drive a turbine (Fig. 42) and the brine sent for well reinjection [35].

A classic example of a wellhead arrangement showing the valves and separator is given in Fig. 43.



Geothermal Power Conversion Technology. Figure 40

A double-flow turbine rotor and cross section of a double-flow dual-admission steam turbine (Courtesy of ICE Costa Rica)



Geothermal Power Conversion Technology. Figure 41
Simplified scheme of a dry-steam power station

The steam from the turbine is condensed by means of either a surface-type condenser (C), as shown in Figs. 16a and 42, or in a direct-contact condenser of either the barometric or low-level type, Fig. 16b.

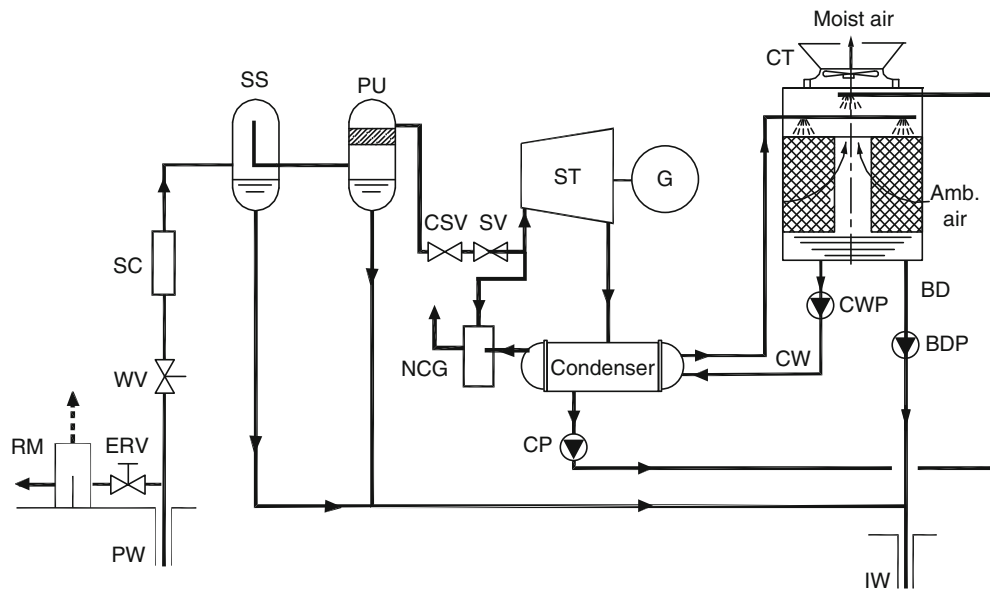
Turbines for Double-Flash Steam The double-flash steam power station is an improvement on the single-flash design as it can produce 15–25% more power output for the same geothermal fluid conditions. The station is more complex, more costly and requires more maintenance. However, the extra power output often justifies the installation of such stations. According to DiPippo [33], 14% of all geothermal stations are double-flash units as of mid-2007.

Many aspects of a double-flash station are similar to a single-flash station. The fundamental new feature is a second flash process imposed on the separated liquid

leaving the primary separator to generate additional steam at a lower pressure than the primary steam.

A schematic diagram of a double-flash station is shown in Fig. 14 [35] and in Fig. 44. The design differs from the single-flash station in Fig. 42 in that the low pressure steam from an additional flasher F flows through a steam line to the turbine in addition to the high-pressure line from the separator.

The turbine must be a dual-admission, single-flow machine to accommodate the high- and low-pressure steam supplies. The low-pressure steam is admitted to the steam path at an appropriate stage to smoothly integrate with the partially expanded high-pressure steam. Other designs are possible e.g., two separate turbines could be used, one for the high-pressure steam and one for the low-pressure steam. In such cases, the turbines could exhaust to a common



PW - Production Well
IW - Injection Well
PU - Purifier
CSV - Control Valve
RM - Rock Muffler
WV - Wellhead Valve
SC - Scrubber
RV - Release Valve
SS - Steam Separator

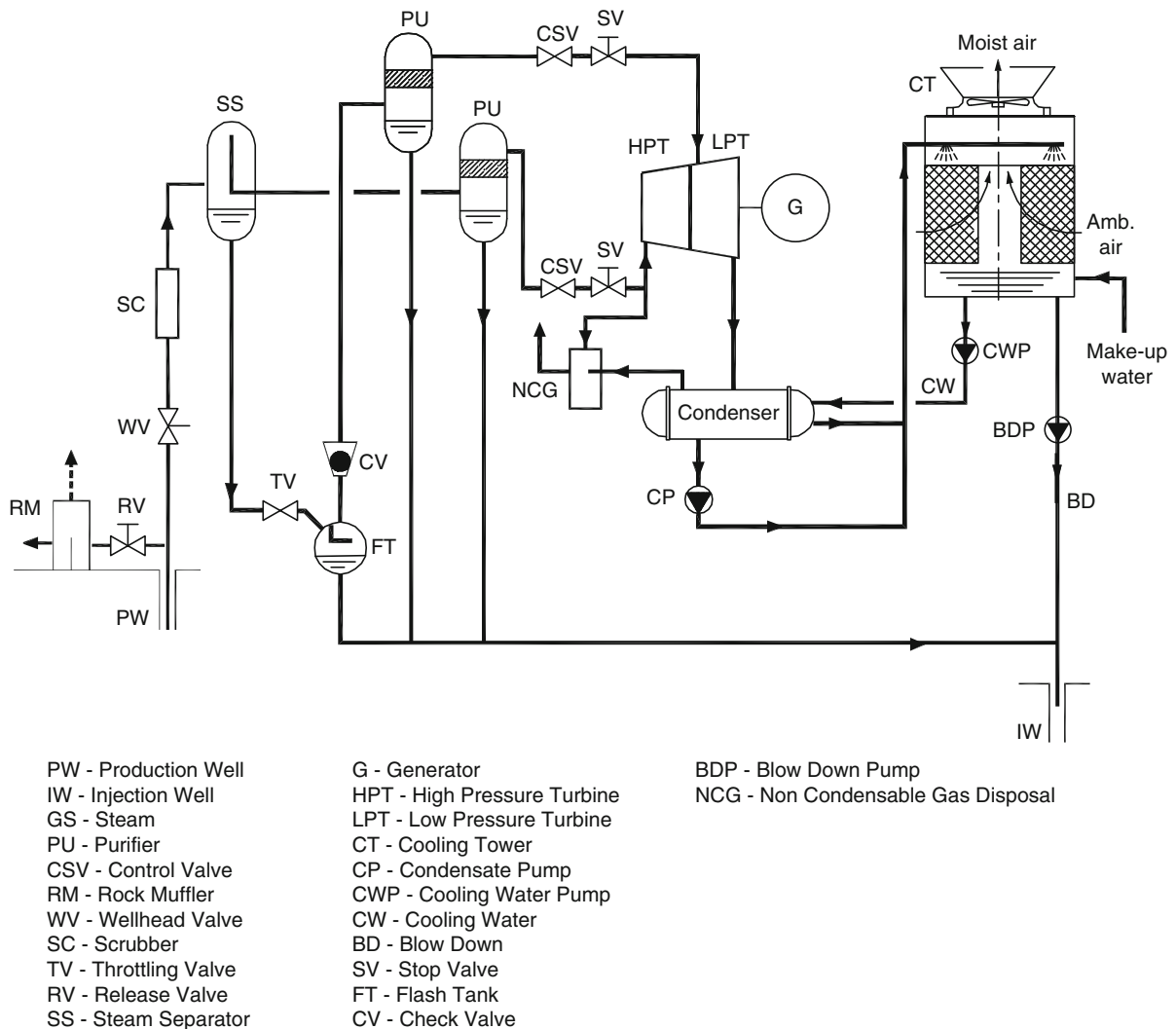
G - Generator
ST - Steam turbine
CT - Cooling Tower
CP - Condensate Pump
CWP - Cooling Water Pump
CW - Cooling Water
BD - Blow Down
SV - Stop Valve

BDP - Blow Down Pump
NCG - Non Condensable Gas Disposal

Geothermal Power Conversion Technology. Figure 42
Simplified single-flash power station schematic [56]



Geothermal Power Conversion Technology. Figure 43
Wellhead valve, control valve (left), and vertical separators (right) (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 44
Simplified double-flash power station schematic [35]

condenser or to two separate condensers operating at the same or different levels of vacuum. For larger power ratings (50 MW or higher), double-flow turbines are preferred to minimize the height of the last stage blades.

Separators and Purifiers

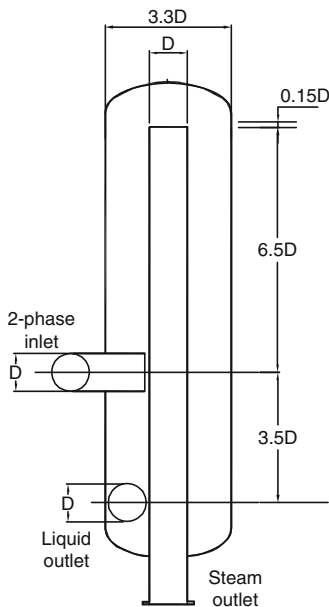
Particle Separators Particles separators/purifiers are installed on steam lines after the wellhead flow control valve to remove particulates carried by the steam flow. The design is based on filtration by circular flow. Restrictions that create differences in

flow velocity help the filtering of the steam-carried particles. To maintain correct separator velocity, more than one separator is installed on a single steam line, see Fig. 45. The collected brine and particles are collected from the bottom exit of the separator.

Separators It is important to separate the two phases efficiently prior to admitting the steam to the turbine. Liquid in the steam can cause scaling and/or erosion of piping and turbine components. Although there are a few designs in use for cyclone separators, the industry has generally settled on the simple Weber-type separator, depicted in Fig. 43 right side and



Geothermal Power Conversion Technology. Figure 45
Particle separators/purifiers (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 46
Scheme and photo of steamboat hills two-phase vertical cyclone separators and accumulators (Courtesy of ORMAT)

Fig. 46. Lazalde-Crabtree [54] published an approach to designing such vessels. The paper presented two variations. One for a primary 1-phase separator and the other for a moisture remover. Their recommended

guidelines for separators and moisture removers are summarized in [Table 7](#).

In cases where the separators are situated at a distance from the power building, the steam

Geothermal Power Conversion Technology. Table 7
Separator and moisture remover design guidelines [54]

Parameter	Separator	Moisture remover
Maximum steam velocity at the 2-phase inlet pipe	45 m/s	60 m/s
Recommended range of steam velocity at the 2-phase inlet pipe	25–40 m/s	35–50 m/s
Maximum upward annular steam velocity inside cyclone	4.5 m/s	6 m/s
Recommended range of upward annular steam velocity inside cyclone	2.5–4 m/s	1.2–4 m/s

transmission pipelines are fitted with steam traps (ST) to capture and remove moisture that may form from condensation within the pipes. Prior to being admitted to the turbine, the steam may be scrubbed to remove any fine moisture droplets that may have formed in the transmission pipelines and escaped the steam traps. The moisture remover that is also known as purifier (PU) is usually located directly outside the power building.

Cyclone separators receive incoming two-phase flow via a tangent entry. The incoming stream circulates tangentially at high speed, so that the liquid flows down the separator inner walls and exits to an accumulating tank. The steam (still rotating) remains in the cyclone separator losing small droplets which exit through a tube on the lower side of the separator, while the steam exits through exterior pipes in the center of the separator. From the accumulator(s) the brine is collected and sent to the reinjection well.

Purifiers Purifiers are needed to protect the turbines blades and structure from corrosion and impingement by removal of residue water droplets. Droplets are either carried by the flowing steam after separation (separation is not 100%), or formed by unintentional condensation on the steam pipe walls. Lazalde and Crabtree [54] recommend a cyclone design similar to the separator but with small water exit due to the smaller water flow load. The basic design parameters are presented in Table 7.

Steam Scrubbing Unit Another method of final steam treatment is scrubbing. To improve the scrubbing,

wash water is injected into the pipeline. TDS concentration in the pipeline is reduced by mixing the low TDS wash water with the high TDS brine drops. Refer Fig. 47.

Steam line scrubbing is effective in removing liquid from the steam while maintaining a low TDS concentration in the liquid drops entering the turbine as steam impurities. The water injection system is shown in Fig. 48.

Flash Tank/Flash Chamber Flash chambers are vertical (Fig. 49) or horizontal (Fig. 50) with brine exits at the bottom of the tank. In Brady Power station, vapors from the first flash pass to two separate steam turbines while the brine that goes to the second flash chamber is re-flashed with its flash steam passing to a third steam turbine. A set of silencers is installed behind the flashing chambers in the case of a trip for part or all of the system.

Flash chambers may be installed horizontally above the condenser and vacuum system. The horizontal chamber has a large volume allowing droplets to settle as a result of low steam velocity. This reduces pressure losses.

Condensers

Surface Condensers In the use of surface-type condenser shown in Figs. 51 and 52, the required flow rate of cooling water \dot{m}_{cw} related to the steam flow rate $X_2 \dot{m}_{st}$. This is expressed by the First Law of thermodynamics as:

$$\dot{m}_{cw} = X_2 \dot{m}_{st} \left[\frac{h_5 - h_6}{\bar{c} \Delta T} \right] \quad (72)$$

where \bar{c} is the assumed constant specific heat of the cooling water (4.2 kJ/kg.K), ΔT is the rise in cooling water temperature at the condenser inlet and outlet and X_2 is the steam dryness stage at the turbine exit.

Direct Contact Condensers For a direct-contact condenser (Fig. 53), the suitable equation is:

$$\dot{m}_{cw} = x_2 \dot{m}_{total} \left[\frac{h_5 - h_6}{\bar{c} (T_6 - T_{cw})} \right] \quad (73)$$

The condenser is below the flash chamber, see Fig. 54. The steam returning from the steam turbine is condensed via direct contact with the cooling water.



Geothermal Power Conversion Technology. Figure 47
Amatatilan steam scrubber, horizontal mixer, and vertical purifier (Courtesy of ORMAT)

Because of the open cooling cycle, air and other gases are entrained in the water. The NCG system must remove this air and gas to maintain the designed condensing temperature/pressure.

Air-Cooled Condensers Air-cooled condensers came to solve the problem of water scarcity in many geothermal sites, as well as to answer the environmental ruling against cooling towers plumes, etc. Air-cooled condensers are rarely used with flash steam mainly as they suffer from internal silica buildup. Today they are used in binary cycles despite the fact that the condensing temperature is higher by 15°C with water cooled condensers, but in arid zones it is the only solution. Air-cooled condensers are very sensitive to ambient air

(temperature), and it can be controlled to some extent by fan speed control, higher exit shroud and optional water spray into the incoming air at extreme air temperatures. HTRI [55] design and similar software is commonly used for air-cooled condensers design. See Fig. 55 scheme and photo of actual air condensers arrangement.

$$\dot{m}_{\text{air}}(h_{\text{airout}} - h_{\text{airin}}) = \dot{m}_{\text{wf}}(h_2 - h_3) \quad (74)$$

where h_2 is organic fluid condition at the condenser inlet and h_3 is the condition of the condensed liquid.

Assuming constant C_p for air in the relevant temperature range and neglecting humidity influence:

$$\dot{m}_{\text{air}} C_{p\text{air}}(T_{\text{airout}} - T_{\text{airin}}) = \dot{m}_{\text{wf}}(h_2 - h_3) \quad (75)$$

or

$$\dot{m}_{\text{air}} = \dot{m}_{\text{wf}} \frac{(h_2 - h_3)}{C_{p,\text{air}}(T_{\text{airout}} - T_{\text{airin}})} \quad (76)$$

Cooling Tower The cooling water used for heat rejection in the condenser of steam or binary stations is

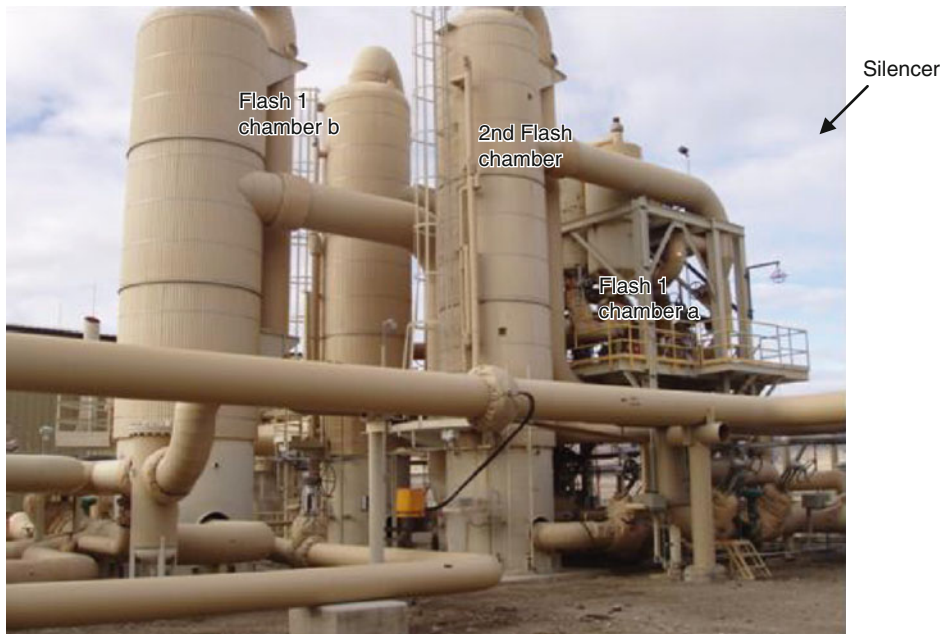


Geothermal Power Conversion Technology. Figure 48
Mokai 2 – Brine scrubbing on steam line (Courtesy of ORMAT)

continuously re-cooled in the cooling tower. Cooling is achieved by evaporation of part of the circulated water into the ambient air. In large cooling towers, cooling can be done by natural draft structure while in medium power systems (in geothermal stations), towers usually are equipped with mechanical draft fans.

The usual temperature difference in mechanical draft towers is about 10°C between the inlet and outlet of the moving air. Cooling water can reach about 25°C depending on the ambient air temperature and humidity. Cooling towers need makeup water to compensate for the evaporation and drift losses and to maintain water quality. The makeup is about 3% of the circulated water depending on ambient air characteristics. In geothermal flash power stations, there is sufficient water quantity collected from the flash steam condensate. In such stations, the condenser is usually of direct contact design which is simpler to design and is more cost effective in production than surface condensers used in binary stations (Fig. 56).

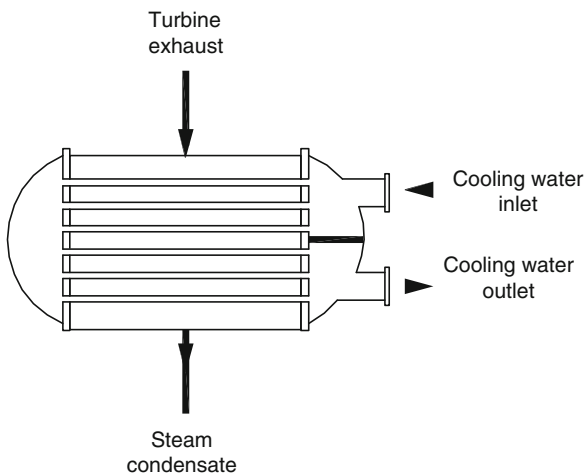
The internal process involves the exchange of both heat and mass between air and water. The following First Law equation describes the overall tower operation, excluding the fan while assuming steady flow and



Geothermal Power Conversion Technology. Figure 49
Brady Double-Flash chambers with back silencers (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 50
GEM horizontal flash chamber (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 51
Surface condenser

overall adiabatic conditions for a tower with direct contact condenser:

$$\dot{m}_1 h_1 - \dot{m}_2 h_2 = \dot{m}_{Aout} h_{Aout} - \dot{m}_{Ain} h_{Ain} + \dot{m}_{BD} h_{BD} \quad (77)$$

$$\dot{m}_1 + \dot{m}_{WAin} = \dot{m}_2 + \dot{m}_{BD} + \dot{m}_{WAout} \quad (78)$$

(Conservation of water)

$$\dot{m}_{Aout} = \dot{m}_{Ain} \quad \text{(Conservation of dry air)} \quad (79)$$

where the terms \dot{m}_{wa} and \dot{m}_{wd} represent water content of the incoming and leaving air streams, respectively. These contents can be determined from the specific humidity, ω , of the air streams:

$$\dot{m}_{WAin} = \omega_{Ain} \dot{m}_{Ain} \quad (80)$$

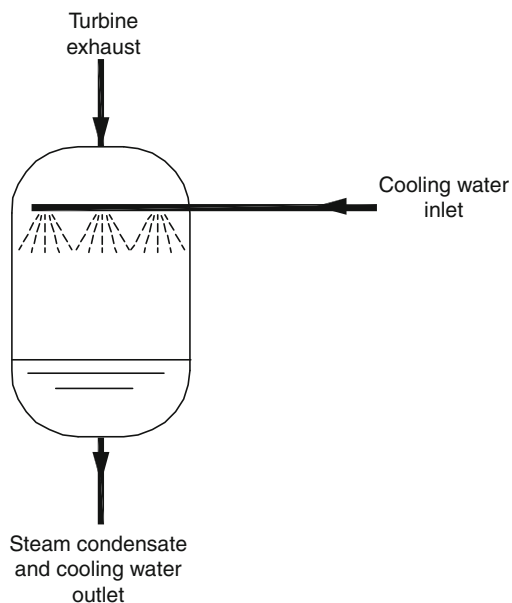
and

$$\dot{m}_{WAout} = \omega_{Aout} \dot{m}_{Aout} \quad (81)$$



Geothermal Power Conversion Technology. Figure 52

Surface condenser end connections, NCG removal system, and cooling tower (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 53

Direct contact condenser

These governing equations are used with the properties of steam, water and moist air, either in tabular, graphic (psychometric chart), electronic form

to determine the various flow rates needed for given design conditions.

Cooling towers are characterized by two parameters:

- Range
- Approach

The range is the change in water temperature as it flows through the tower, namely, $T_1 - T_2$.

The approach is the difference between the water outlet temperature and the wet-bulb temperature of the incoming air, namely, $T_2 - T_{wb.Ain}$. Since the ideal outlet water temperature is the wet-bulb temperature of the incoming air, the approach is a measure of how closely the tower approaches ideal performance, i.e., zero approach or $T_2 = T_{wb.Ain}$ (Fig. 57).

Pumps

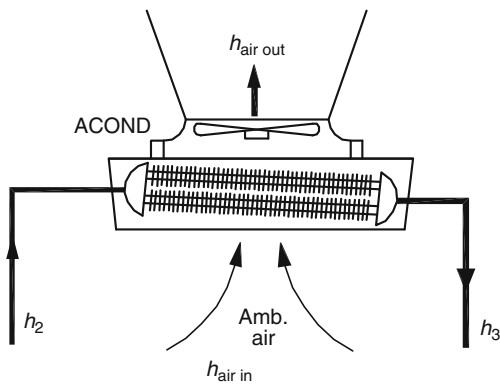
Geothermal Fluid Pumps

Production Pump Production pumps are of multi-stage vertical design. The pump fits the bore diameter of the well casing and has a screen filter before its inlet opening. The motor sits directly on the wellhead, on top of the pump exit pipe. See Fig. 58. Pump pressure



Geothermal Power Conversion Technology. Figure 54

GEM direct contact condenser, hot-well pumps, and vacuum system (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 55

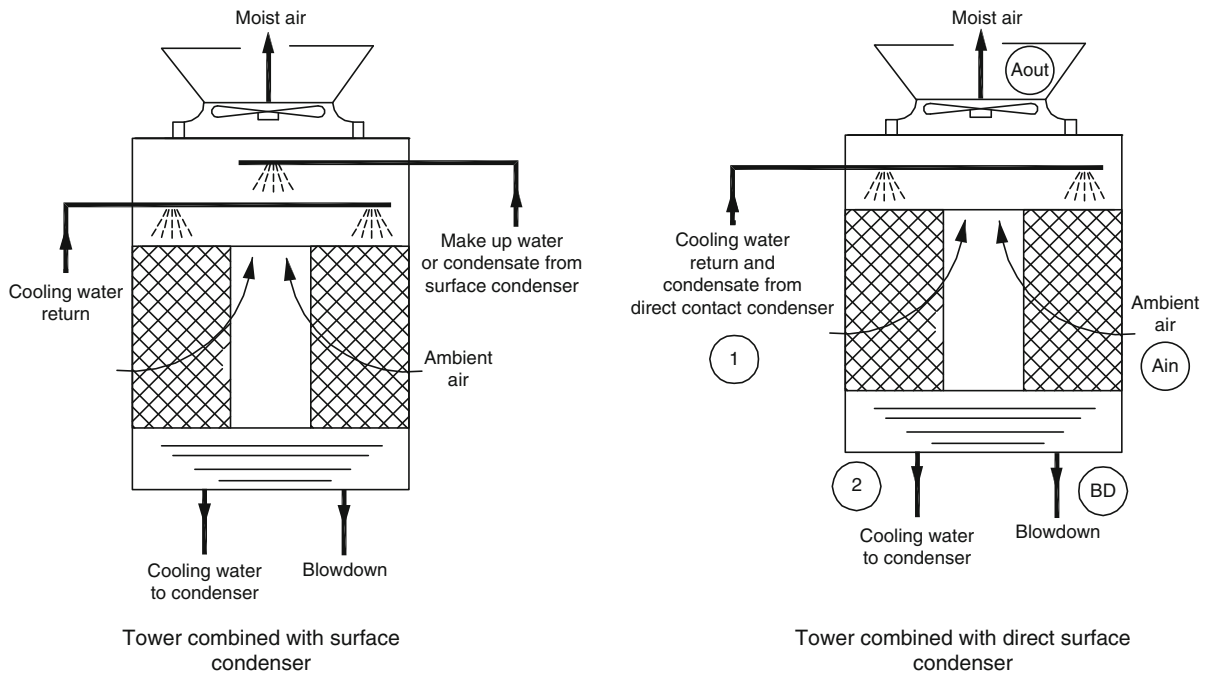
Air-cooled condenser view of the PGV Geothermal combined cycle power station (Courtesy of ORMAT)

should be high enough to overcome all system piping friction while still maintaining pressure above possible precipitation of carbonates as mentioned in section on “[Geothermal Resources](#).”

Injection Pump The injection pumps are of horizontal or vertical multistage design to overcome injection well

resistance. See [Fig. 59](#) for typical arrangement. The structure must allow quick dismantling for maintenance work.

Condensate Pumps Condensate pumps overcome the subatmospheric pressure at the pump inlet and then lift condensate to the top of the cooling tower, overcoming



Geothermal Power Conversion Technology. Figure 56
Induced draft cooling towers showing related mass flows



Geothermal Power Conversion Technology. Figure 57
Cooling towers of GEM single-flash power station in East Mesa, California (Courtesy of ORMAT)

distribution system and nozzles resistance. Typical pump is in Fig. 60.

Binary Station Motive Fluid Pump Motive fluid pump is situated below the condenser structure. To eliminate possible inlet vacuum buildup it is usually a vertical

with barometric height above its inlet. Typical binary cycle circulation pump is in Fig. 61.

Gathering System

Dry-Steam Gathering The connection between the wells and the power building for a dry-steam power



Geothermal Power Conversion Technology. Figure 58
North Brawley production pump and separator with MCC (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 59
Injection pumps arrangement (Courtesy of ORMAT)

station is relatively simple. At the well, there are the usual valves and a steam purifier. See example in [Fig. 62](#). The purifier is usually an in-line, axial

centrifugal separator designed to remove all carried particles from the steam before it enters the piping system. Steam pipes are insulated and include high expansion loops. See example in [Fig. 64](#). Steam traps are sited along the pipes to remove condensate.

At the wellhead, or immediately before the steam approaches the power building, there is an emergency pressure relief valve. This allows for the temporary release of steam in the event of a turbine trip. Before being released to the atmosphere, the steam generally passes through a silencer ([Figs. 49](#) and [63](#)). It has been found preferable to maintain the wells in a steady open mode rather than cycling the wells through open and closed positions. At the power building there is a steam header, a final moisture remover (typically a vertical cyclone separator or a baffled demister), and a venturi meter for accurate steam flow rate measurement.

Water and Brine Piping High-pressure reinjection requirements might require pumps to maintain sufficient injection pressure.

Pressure drop through the pipes due to friction and local losses can be calculated using standard handbooks such as McKetta or similar [[56](#)]. The information required includes fluid mass flow rate, fluid density,



Geothermal Power Conversion Technology. Figure 60
Condensate pump (Courtesy of ORMAT)

fluid dynamic and kinematic viscosity, pipe data such as pipe diameter, friction factor and length.

If there is a change in pipe elevation, the gravity contribution must be included.

Pressure loss in a two-phase, steam-liquid pipeline is more complex and less reliable for analytical prediction [57, 58]. Correlations may be used to establish the pressure drop. Field tests are conducted to experimentally determine the exact ΔP . The situation is complex as the two-phase flow in any of several different patterns depends on the pipe orientation and relative amounts of the phases present. See also references [58, 59] for suggested calculation depending on type of flow

Steam Piping There are three types of fluids flowing in the geothermal field piping system:

- Steam
- Water/brine
- Two phase (mixture of steam and water)

One of the main gathering system design concerns is the pressure loss in the steam lines from the wellhead to the power building. The steam pressure drop is

a function of the diameter, length, configuration of the steam piping and the density and mass flow rate of the steam. See also Handbooks for fluid systems design such as ref. [56] or similar.

Note that the calculated pressure drop might be higher than the actual values at higher velocities. The central variable is the pipe diameter as pressure drop is inversely proportional to the pipe diameter in the fifth power.

By installing larger diameter pipes pressure loss can be considerably reduced. The extra cost of the larger pipes may be a negative economic factor. A thermodynamic-economic optimization study will determine optimum pipe size.

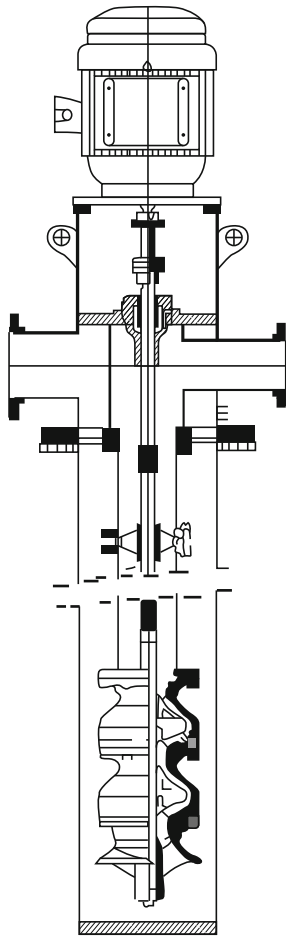
Single-Flash Gathering System Design Considerations and Piping Layouts When a geothermal field produces a mixture of steam and water, the method used for energy conversion depends on the potential energy available in each of the streams. The curves of maximum available energy for steam and water given in Fig. 8 can be used for the initial analysis. If the separated water stream is not sufficient for power generation, it will be reinjected, while the separated steam is utilized in a single-flash station for conversion to electricity. Each power station comprises a number of production and reinjection wells to assure continuous flow even during maintenance work on any of the wells. A piping system is required between the wells and the installed steam/water separators (usually adjacent to the wells and the power building). There is also a brine piping system leading from the separators and power building to the reinjection wells.

A typical field is usually a few kilometers long with the production wells on one side, injection wells on the other side and with the power building positioned so as to minimize steam side pressure losses.

Double-Flash Gathering System Design Considerations In most cases, the second flashing process is performed near to the first steam separator.

The list of possible arrangements is large and therefore the best choice will be determined by thermodynamic and economic analysis taking site-specific conditions including:

- Temperature, pressure, and chemical nature of the geothermal fluid



Geothermal Power Conversion Technology. Figure 61
Binary power station – motive fluid circulation pump (Courtesy of ORMAT)

- Location of production and injection wells relative to the power building
- Topography of the site
- Method of fluid disposal, including any required scale-control techniques

For such analysis the pressure drop calculations for the various piping arrangements can use the formulas in “[Water and Brine Piping](#)” and “[Single-flash Gathering System Design Considerations and Piping Layouts](#).”

Two-Phase Flow The two-phase pipelines can be designed as elements of a geothermal gathering system. Take the correct pressure drop into account as it can be larger than that in single phase steam lines.

The presence of unsteady flow patterns such as slug flow can cause excessive vibrations and should be avoided by proper pipe diameter selections. So-called flow pattern “maps” [57] can help the designer with the correct regimes.

Another important aspect concerns the flow of liquid removed from the cyclone separators. The fluid is in a saturated state and any pressure loss can cause it to flash into vapor. This will create a vapor barrier and inhibit the fluid flow down the well. In such cases it may be necessary to bleed the vapor from the wellhead or to install a booster pump upstream of the vapor breakout point. Also, any drop in temperature of the liquid may change the chemical equilibrium and cause precipitation as described in section on “[Geothermal Resources](#).”



Geothermal Power Conversion Technology. Figure 62

Production wellhead valve and control valves in PGV station, Hawaii (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 63

Rock Muffler in the station at Brady, Nevada (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 64
Brine line with expansion loop (Courtesy of ORMAT)

Choosing the Energy Conversion Systems

Introduction

Four basic types of geothermal energy conversion systems were covered in the section on “[Thermodynamic Analysis of the Energy Conversion Process](#).” There are geothermal resources demanding more sophisticated energy conversion systems than the basics considered until now. Furthermore, energy conversion systems have evolved to fit the needs of specific developing fields by integrating different types of power station into a complex facility described later.

Of the over 10,715 MW of geothermal stations in operation worldwide, most are steam stations operating on dry steam or steam produced by single or double-flash. About 1,000 MW use ORC or steam/ORC combined cycles [60, 61]. [Table 8](#) compares various resources fluids and their temperatures that indicate site potential and recommended configuration.

Operational experience has confirmed the advantages of the ORC stations, not only for the low enthalpy water-dominated resources, but also at high enthalpy with aggressive brine or brine with high noncondensable gas content. The higher installation cost of these systems is often justified by environmental and long-term resource management considerations [62, 63]. The air-cooled ORC stations are particularly well adapted to the engineered geothermal systems (EGS).

Optimization of the Design of the Power Cycle

The optimization of the whole geothermal power station system is accomplished by matching the working cycle and fluid properties to the resource characteristics, when considering not only resulting efficiency and cost, but also the impact on environment, long-term pressure support, requirements for makeup wells and O&M costs.

Resource Considerations Sustainability is defined as the ability to economically maintain the installed capacity over the life of a station [64]. With geothermal power stations, this is controlled by two factors, heat recharge and water recharge.

Sustainable heat flow to the station, beyond the natural heat recharge, is supported by accessing the stored heat through drilling additional wells over the life of the project.

The decline of production in the Larderello, The Geysers, and Wairakei fields has focused attention on the necessity for long-term pressure support by injecting as much of the geothermal fluid as possible back into the aquifer.

In brines rich in carbonates, flashing, as accomplished in conventional steam power stations leads to scaling of injection wells, reducing their life span.

Use of secondary loops and of downhole and booster pumps, as in air-cooled ORC power stations

Geothermal Power Conversion Technology. Table 8 Comparison of basic geothermal energy conversion systems

Resource	Temperature	NCG	Dissolved solids	Configuration
Water	High or medium	Low	Low	Condensing steam (double-flash) or ORC
		High	Low	ORC
		Low	High	ORC
	Low	Any	Any	ORC
Water Dominated	High or Medium	Low	Low	Condensing steam double-flash or single-flash + ORC
		High	Low	ORC
		Low	High	ORC
	Low	Any	Any	Two-phase ORC
Steam Dominated	High or Medium	Low	Low	Condensing steam (single or double-flash) or condensing steam (single-flash) + ORC
		High	Low	Integrated Geothermal Combined Cycle
		Low	High	or Two-phase ORC
	Low	Any	Any	Two-phase ORC
Dry Steam	High or low	Low	Low	Condensing Steam
		High	Low	Geothermal Combined Cycle
		Low	High	Geothermal Combined Cycle
	Very High	Low	Low	Triple Flash Condensing
		High	Low	Geothermal Combined Cycle
		Low	High	Geothermal Combined Cycle

enhance sustainability by assuring complete water recharge while reducing both the fouling of heat exchangers and the scaling of injection wells.

Heat Cycle Considerations When the source is liquid phase only (sensible heat) the ideal cycle would have a varying source temperature of a succession of infinitesimal Carnot cycles. In a subcritical Rankine cycle the constant temperature of the evaporation leads to a loss of energy, but because of the lower vaporization latent heat this drawback is lower than in a steam cycle.

The objective of attaining to the ideal cycle has been aimed at in proposing the super-critical Organic Rankine cycle, the total-flow regenerative cycle, the cascaded Organic Rankine cycle and the Kalina cycle. When dry steam is available, the most effective way is to use the conventional condensing steam cycle.

When the source is a mixture of steam and brine and/or has a high content of noncondensable gases, the most effective utilization of the resource is achieved through a combined geothermal cycle by firstly expanding the steam in a back-pressure steam turbine. The heat of condensation together with the heat of separated brine is then used to drive a bottoming ORC.

To compare the efficiency of the different systems it is necessary to consider the net output of parasites, such as cycle pumps, production pumps, injection pumps, cooling systems and noncondensable gas extraction power consumption [63].

Work Ratio, Parasitic Losses, and Impact on Total Cost These considerations arise in every design of new power station. These are many options to consider, a few examples follow below. While the technical issues are general, the economic decision is site specific and

depends on the type of financing, interest rate, contracted price of electricity, etc.:

- Condensing temperature

The heat source temperature is a given factor that cannot be modified. However, the decision on the condensing temperature is negotiable and is a result of technical and economical considerations. Additional heat exchange area to a surface condenser or air-cooled condenser improves the work ratio, but adds cost of hardware. The question is whether the cost of additional kW installed or additional kWh produced per year can be justified economically.

- Single-flash or double-flash steam cycle

Although the technical advantage of double flashing is understood and is considered to add about 20–30% to the output per well, actual site evaluation is still required. The factors to be considered are source size, heat source temperature, ratio between water and steam, etc. Additional pieces of equipment and different turbines may be required. All this affects the relative cost against the energy gain.

- Subcritical or supercritical operation of binary cycle

Subcritical operation system is simpler to construct and operate but has limitations in heat transfer between the brine and working fluid. Supercritical cycles improve the heat transfer but require high-pressure operation increasing the pump power.

- Cascading vs. supercritical cycle

A cascading binary cycle system is a simple solution to heat transfer improvement, while avoiding supercritical cycle limitations. The gain in power output is considered against cost of equipment. Even though multistaging is theoretically the best for gain in efficiency, the practical cost comparison dictates dual- or triple-stage design. Another consideration in cascading systems is the reduced brine temperature that can cause precipitation in the heat exchangers. This is partially compensated for by keeping the brine side under higher pressure (also has some cost impact).

Cooling System Consideration In steam Rankine cycle systems the use of condensate as makeup water is the most cost effective approach.

In areas with no natural water recharge of the geothermal reservoir and no surface water available, air cooling has been used by implementing air-cooled ORC for low-temperature resources or geothermal combined cycle for high-temperature resources. The value of the air-cooled ORC is important in case of EGS which is highly dependent on water recovery ratio.

Environmental Considerations Use of air-cooled ORC reduces the impact on environment by reinjection of:

- Noncondensable gases (mainly H_2S released by the steam)
- Discharged fluids such as the separated brine (carrying heavy metals) and blow-down from the cooling tower (chemicals) and drift from cooling towers.

Commercial Power Stations

Steam Power Stations

Dry-Steam Power Station

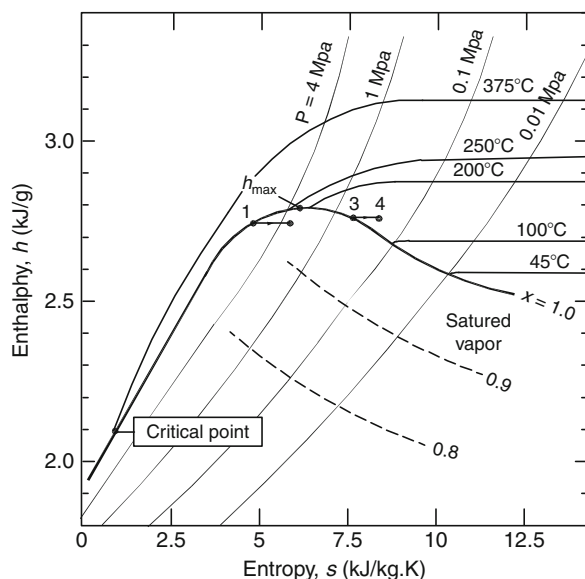
General Dry-steam stations were the first type of geothermal power station to achieve commercial status with their history going back 100 years [66].

Dry-steam stations tend to be simpler and less expensive than flash-steam stations in that there is no geothermal brine to contend with. As can be seen, this is a positive issue when it comes to maintaining reservoir performance.

Large dry-steam reservoirs have been discovered only in two areas of the world; Larderello and The Geysers. There are limited dry-steam areas in Japan (Matsukawa), Indonesia (Kamojang), New Zealand (Poihipi Road section of Wairakei) and the USA (The Geysers, California). White [27] estimated that only about 5% of all hydrothermal systems with temperatures greater than 200°C are of the dry-steam type.

The general characteristic of a dry-steam reservoir is that it comprises porous rocks featuring fissures or fractures, either occluded or interconnected, that are filled with steam. Whereas the steam also contains gases such as carbon dioxide, hydrogen sulfide, methane, and others in trace amounts, there is little or no liquid present.

The dry steam extracted from the mentioned resources is either saturated or slightly superheated at



Geothermal Power Conversion Technology. Figure 65 Mollier chart for water (maximum enthalpy at $T \sim 235^\circ\text{C}$ and $P \sim 3.07\text{ MPa}$)

temperatures near 235°C and pressure near the maximum saturation in Molier curve (30.7 bars) as in Fig. 65. Isenthalpic pressure loss in the upper layers (1–2, or 3–4, respectively) explains the superheated condition at the turbine inlet, but does not explain how the steam sometimes remains saturated at the wellhead.

There are over 60 flash or dry-steam commercial stations in operation, each with average power of 40 MW.

Energy Conversion System Once the steam reaches the power building, a dry-steam station is essentially the same as a regular low-temperature boiler steam station. The turbines are single-pressure units with impulse-reaction blading, either single-flow for smaller units or double-flow for large units above 50 MW. The condensers can be either direct-contact (barometric or low-level) or surface-type (shell-and-tube). For small units it is often advantageous to arrange the turbine and condenser side-by-side for maintenance reasons.

A typical dry-steam power station scheme is shown in Fig. 41 and the corresponding steam process is in Fig. 13 (repeated here in Figs. 66 and 67). Since the wells produce saturated steam (or slightly superheated

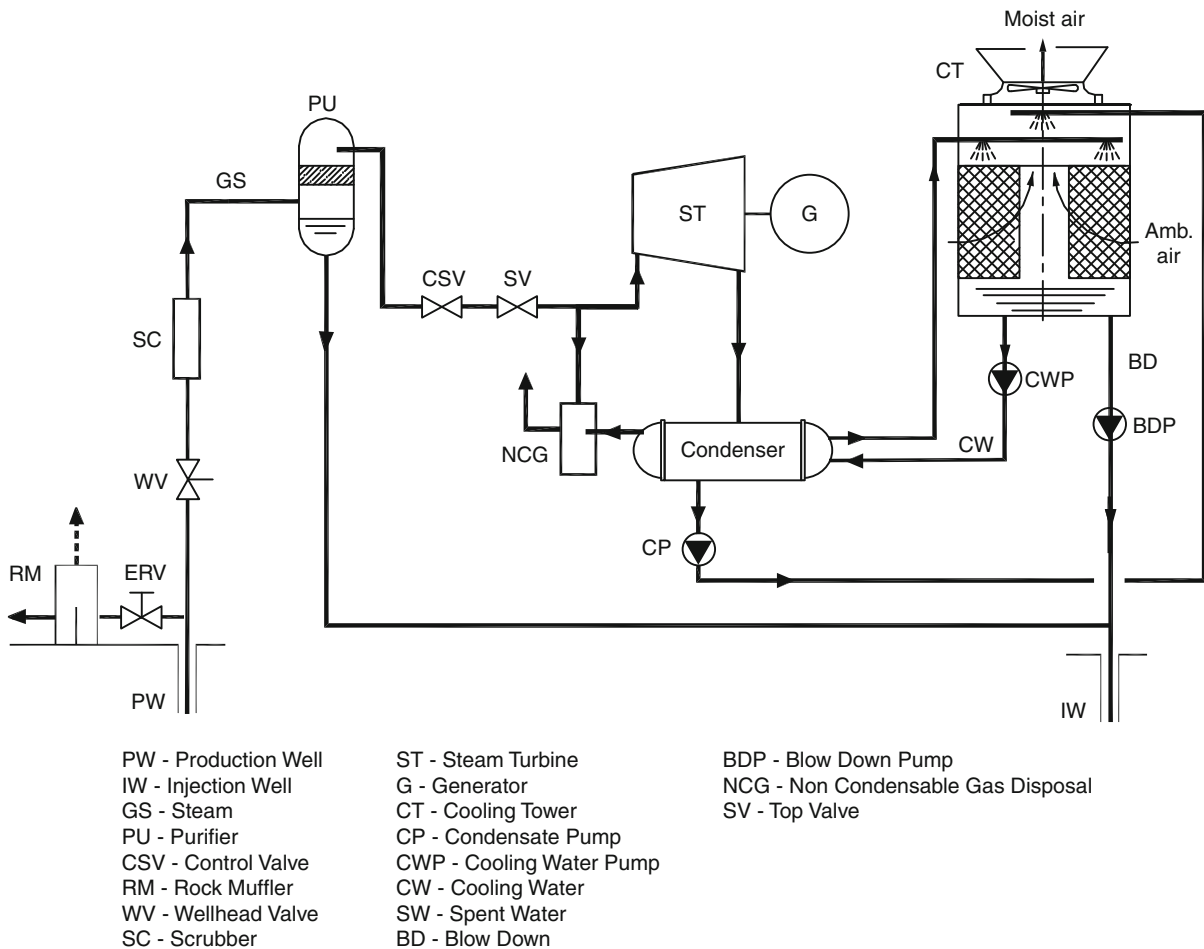
steam), the starting point (state 1) is located on the saturated vapor curve. The turbine expansion process 1–2 generates somewhat less power output than the ideal, isentropic process 1–2s. Heat is discarded to the surroundings in the condenser via the cooling water in process 2–3.

The turbines used in geothermal applications must be made of corrosion-resistant materials owing to the presence of gases such as hydrogen sulfide that can damage ordinary steel.

Single-Flash Power Stations Since single-flash stations have a significant amount of waste liquid from their separators that is still fairly hot (typically $150\text{--}170^\circ\text{C}$), this can be used to generate more power instead of being directly injected. Combined single- and double-flash stations have been built at several fields around the world.

Double- and Triple-Flash Power Stations Double-flash cycles as in Fig. 44 are justified due to the high temperature of the waste brine remaining from the first flash. For such cases the turbines are designed to handle dual-pressure steam. Also, to maintain symmetric axial force on the turbine bearings and shorter blade height they are designed as double-flow machines. See scheme in Fig. 68. In some cases the source temperature and flow-rate justifies triple flashing as in the case of NGA AWA PURA power station in New Zealand shown in Fig. 69. The general station scheme is given in Fig. 70 and the turbine steam flow is given in Fig. 71. The Fuji 120 MW turbine is given in Fig. 72. However, mass flow does not always justify the construction of the special two admittance turbine. This lead to combinations as integrated single and double-flash stations, combined single and double-flash units as will be described later.

Integrated Single and Double-Flash Stations Single-flash units have been built and have been operating for a period of time where the geothermal fluid reservoir temperature is about $220\text{--}240^\circ\text{C}$. The addition of one more flash using the separated brine allows for a lower pressure unit. The arrangement shown in Fig. 73 consists of two single-flash units, Units 1, Unit 2 and Unit 3, added at a later date appears to be simply another single-flash unit. The power station as a whole is an integrated single- and double-flash facility since the



Geothermal Power Conversion Technology. Figure 66
Simplified scheme of a dry-steam power station

original geothermal fluid experiences two stages of flashing [35].

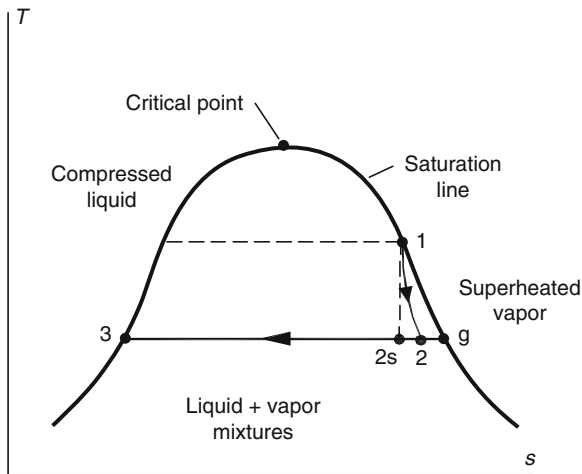
The advantage to this arrangement is that no new wells need to be drilled to supply the third unit. Unit 3 serves as a bottoming unit recovering some of the wasted potential from the still-hot brine. The thermodynamic process diagram is given in Fig. 74.

One possible thermodynamic drawback to this arrangement lies in the selection of the pressure (or equivalently, the temperature) for the second flash process 3–6 in Fig. 74. If the flash temperatures for the first two units (assumed identical) had been optimized, then the addition of the third unit requires not only new optimization but evaluation of possible use of the existing equipment.

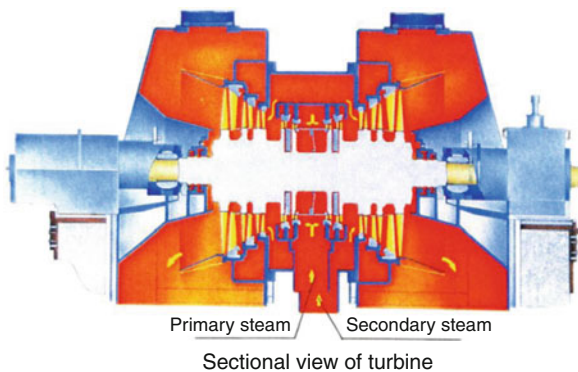
Combined Single- and Double-Flash Stations

When the resource temperature is equal to or greater than 240°C, it may be possible to augment the single-flash units with a true double-flash bottoming cycle, as in the schematic flow diagram Fig. 75, and in the process diagram Fig. 76. For this case, the waste brine from the first units is subjected to two flashes, resulting in two additional low-pressure steam flows to be utilized in a dual admission-pressure turbine. The arrangement can be named as “combined single- and double-flash station” [35].

Although the thermodynamics of this arrangement are favorable, i.e., a higher resource utilization efficiency than for the original single-flash station, there may be problems with chemical scaling. This due to



Geothermal Power Conversion Technology. Figure 67 Temperature-entropy diagram for dry-steam station (steam saturated at the turbine inlet)



Geothermal Power Conversion Technology. Figure 68 Scheme of double-flow dual-pressure turbine (Courtesy of Mitsubishi)

silica precipitation at low temperatures associated with the last flash as discussed in section on “[Geothermal Resources](#).” Therefore few flashes would not be a good choice unless there is no possibility of silica precipitation or if silica treatment is considered as part of the investment.

Power Stations Using Hyper Saline Brines In some geothermal sites, the underground soil formation is comprised of various materials that when dissolved

into the hot water cause the water to become acidic and saline. This water can clog a production well or the downstream piping and heat exchangers in a few days, or render surface vessels useless by contamination.

One of the most notorious geothermal resources is located in the Imperial Valley of southern California, near the southeastern shore of the Salton Sea. The resource was recognized in the 1850s when explorers moving west came upon hot pools and mud volcanoes in an otherwise barren desert [67].

Drilling for power production began in the 1960s but the early wells were all plugged and abandoned. Some wells drilled in the 1970s are still in operation today, but the fluids that were produced resisted exploitation for power generation because of severe scaling and corrosion problems. The temperatures were high (up to 360°C), and the total dissolved solids reached as much as 300,000 ppm, placing these fluids in the hyper saline category. The chemical analysis of the fluid produced from the Magmamax No. 1 well drilled in 1972 showed that chlorides, sodium, calcium, and potassium make up about two thirds of the 300,000 TDS in the brine [68].

An extensive research effort began in the 1970s funded by the US Department of Energy, the Electric Power Research Institute and several private companies. Through this effort, techniques were devised that later permitted these fluids to be used for the generation of electricity in a reliable and cost-effective manner [69]. Two approaches for dealing with these aggressive brines have been used with reasonable success, flash-crystallizer, reactor-clarifier (FCRC) and pH modification (pH-Mod) systems. The principles underlying these two methods are detailed below.

Flash-Crystallizer/Reactor-Clarifier (FCRC) Systems In the FCRC approach, clean steam is generated in a train of separators and flash vessels, similar to standard flash-steam power stations, but the separated brine is seeded with material inducing precipitation. A simplified schematic of an FCRC power station is shown in Fig. 77 [69]. The seed material is obtained from the highly concentrated brine waste stream. In this way, the unstable, supersaturated solids precipitate on the seed particles, rather than on the surfaces of the vessels and piping. The particulate matter eventually



Geothermal Power Conversion Technology. Figure 69
Awa Pura Power station in New Zealand (Courtesy of Fuji Electric)

settles in a reactor-clarifier vessel. The slurry from the reactor-clarifier is thickened and a portion of it is recirculated as seed material. The clarified liquid is pumped to a secondary clarifier and then sent to reinjection wells.

The most recent power station to use this approach, Salton Sea Unit 5, has a triple-pressure turbine that receives the high-pressure steam separated at the well-head separators and expands it through the first four stages of the turbine [70]. This eliminates the throttling loss from the pressure-letdown throttle valve TV shown in Fig. 77.

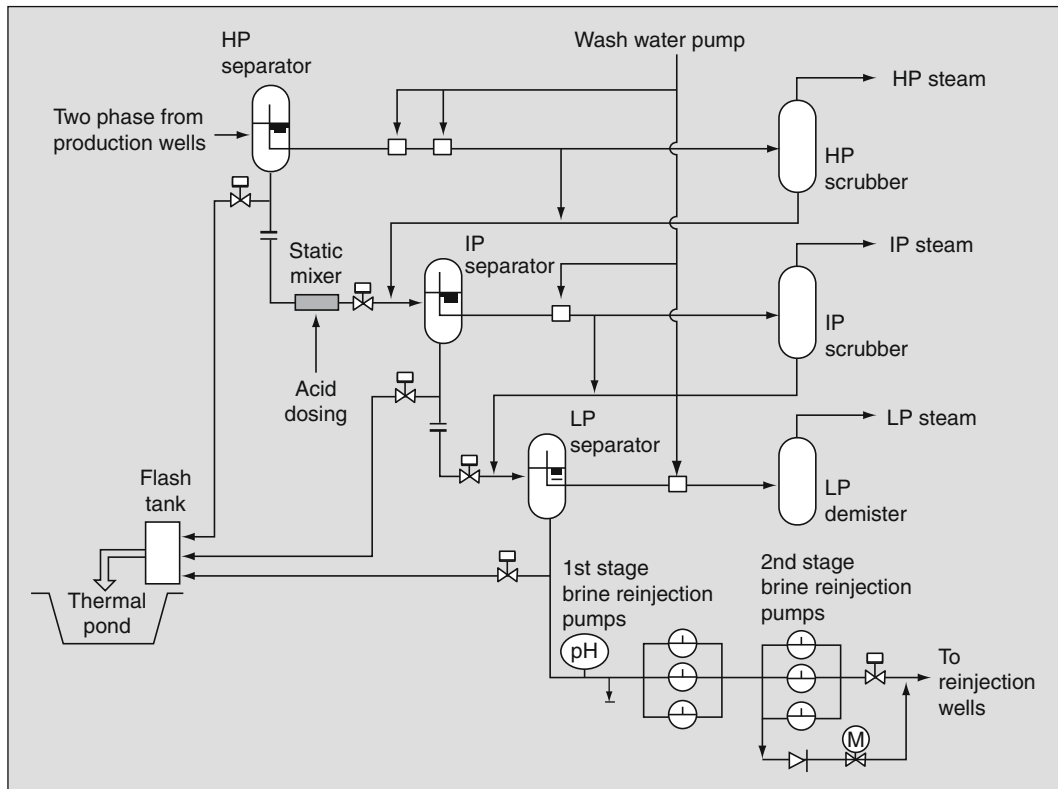
pH Modification (pH-Mod) Systems Section on “pH” discussed the option of pH control to protect against silica scaling. One approach is to modify the brine pH altering the kinetics of the precipitation process. The technique of acidifying the brine has been used in some Salton Sea stations as an alternative to the FCRC approach. Acids proposed for pH control are Hydrochloric acid HCl [71] and various sulfur-based acids [72]. By reducing the pH of the geothermal fluid the solubility of silica is increased, kinetics of the reaction are slowed and it is possible to avert

precipitation, at least until the separated liquid has been processed to generate the flash steam needed for the turbine. A highly simplified flow diagram for a pH-Mod station is shown in Fig. 78 [73].

The addition of hydrochloric acid to the brine requires appropriate corrosion-resistant materials. However, pH-Mod stations are much simpler than FCRC stations in terms of the number of vessels needed and the operating procedures to be followed.

The processing of the discharge waste brine D is omitted from Fig. 78 but treatment may be needed. If the injection faces problems in the pipelines or the injection wells, then the brine pH must be raised. In addition, a reactor-clarifier can be used to remove the silica and assure that the waste brine can be safely injected. Work has been done on the economics of recovered minerals from the brine – which in itself is a by-product [74].

With these two methods for handling Salton Sea brines (FCRC and pH-Mod), it has been possible to construct ten power stations with a total capacity currently generating 327 MW net [75]. It is believed that a significant portion of the ultimate potential of the geothermal field lying offshore beneath the Salton Sea



Geothermal Power Conversion Technology. Figure 70

Nga Awa Pura steam separation system overview (Courtesy of Fuji Electric)

is far from being fully exploited. The technologies described in this section should allow this valuable resource to reach its full potential.

Organic Rankine Cycle Configurations for Geothermal Power Stations

Binary Power Stations

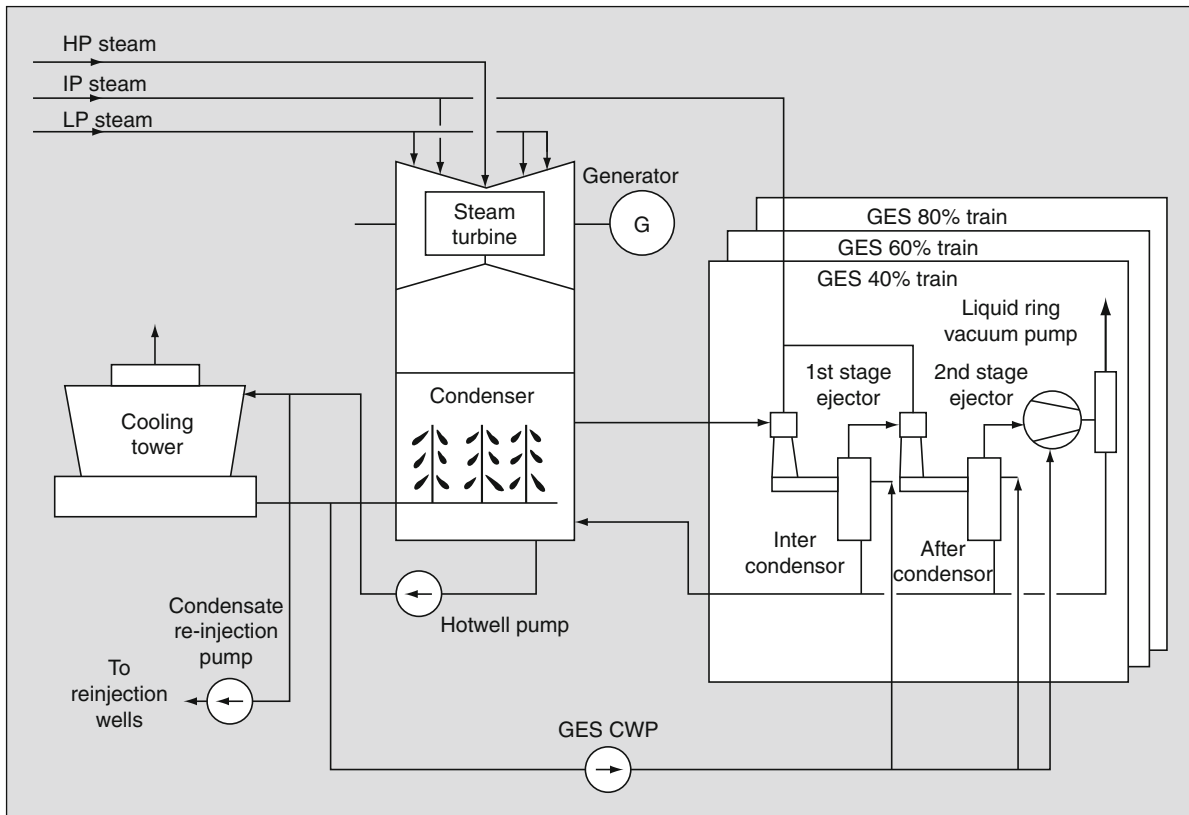
Single-Phase ORC In cases where the geothermal fluid temperature is 150°C or less, flash systems require intensive engineering work including large flashing vessels and brine treatment both in the production and reinjection pipes and wells. In addition, in most cases, at such temperatures a production pump is required to maintain continuous geofluid flow and pressure to prevent scaling.

Although the GEM station at East Mesa in the Imperial Valley of California in the USA [34] flashes the compressed liquid as in Fig. 79, it is simpler to pass the geothermal fluid as a compressed liquid through

heat exchangers and dispose of it (in the liquid phase) into reinjection wells. By improving the heat transfer across the heat exchangers an economically viable design is obtained. A water-cooled, brine-driven binary cycle is in Fig. 80 [35] and an air-cooled system is given in Fig. 81 [61].

The production wells (PW) are fitted with deep well pumps (P) and are set below flash depth determined by the reservoir properties and the desired flow rate. Sand filters SF are used in most cases to prevent scouring and erosion of the piping and heat exchanger tubes. Typically there are two steps in the process. First the working fluid flows through a preheater PH where it is brought to boiling point, then it flows to the evaporator E acquiring the supplemental heat of evaporation, emerging as a saturated vapor. The working fluid then expands in the turbine, recondenses in the condenser, and is pumped back to the preheater via a feed pump.

The geofluid flows to the evaporator, then the preheater and to the reinjection well. The geothermal fluid is



Geothermal Power Conversion Technology. Figure 71

Nga Awa Pura power generation facility overview (Courtesy of Fuji Electric)

always kept at a pressure above its flash point for the fluid temperature to prevent the breakout of steam and noncondensable gases that leads to calcite scaling in the piping. Furthermore, the fluid temperature is not allowed to drop to the point where silica scaling becomes an issue in the preheater, piping, and reinjection wells. Therefore chemical problems mentioned in section on “[Geothermal Resources](#)” are eliminated.

Temperature Cascading Organic Rankine Cycle To increase the power output of a binary power station, a cascading system can be used. In a simple cascading method there are two or more evaporators and preheaters, arranged consecutively in consequent structure. The geothermal fluid travels from one pair of units to the next. The station schematic given in [Fig. 27](#) incorporates three levels of organic systems, each working at a different range of temperatures.

All preheaters begin from the same temperature but evaporation is performed at three different temperatures, therefore three turbines are required for such operation, ensuring the cooled brine is better utilized by the preheaters with part of the preheating in the evaporators. A T-Q diagram of the schematic arrangement of this system is given in [Fig. 28](#).

A two-level cascading based on two turbine integrated over one generator was used in Ormat Heber brine station as given in [Fig. 82](#). The brine enters level 1 evaporator then the level 2 evaporator. From this level the brine goes to the preheaters of both levels. The turbines of both levels are connected to single generator. The turbines of each level can be divided into HP and LP turbines, with each pair driving a single generator. Both condensers are water cooled sharing the same cooling water supply line. The station is shown in [Fig. 83](#).



Geothermal Power Conversion Technology. Figure 72
Nga Awa Pura view over turbine generator (Courtesy of Fuji Electric)

Recuperated Organic Rankine Cycle In most actual cases, the perfect match above is not feasible because of limitation of the brine and condensate mixture cooling temperature. In most of the cases the limiting factor is the silica scaling risk, which increases as the brine temperature drops. A method to partially overcome the cooling temperature limit is to add a recuperator which provides some of the preheating heat from the vapor exiting the turbine.

The recuperator is applicable when the organic fluid is of the “dry expansion” type, a fluid where the expansion in the turbine is done in the dry superheated zone and the expanded vapor contains heat that has to be extracted prior to the condensing stage (see [36, 77]). The recuperated Organic Rankine cycle is typically 10–15% more efficient than the simple Organic Rankine cycle described in the beginning of this chapter (Fig. 66 for comparison). This applies to the two-phase geothermal power station in Fig. 84 with its T-Q diagram in Fig. 85.

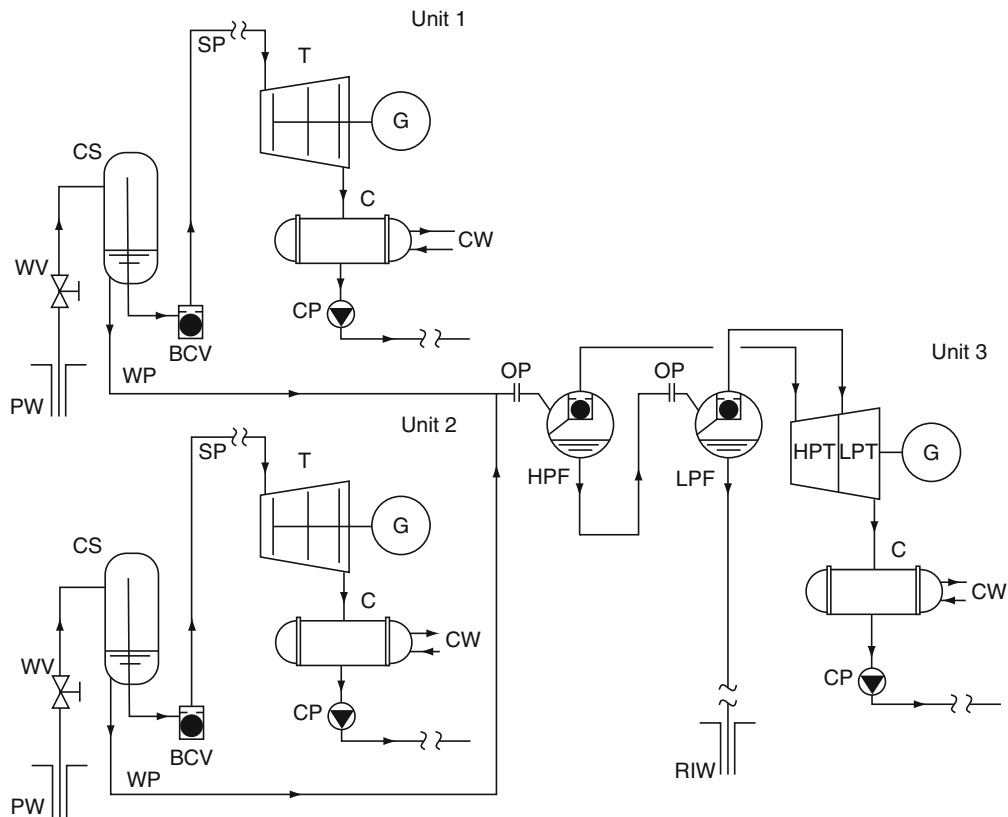
The recuperated two-phase process is used by Ormat in many geothermal projects around the world, i.e., 20 MW Zunil in Guatemala, 14 MW

Ribeira Grande I and II in San Miguel in the Azores (Fig. 86), and 1.8 MW Oserian and 13 MW Olkaria III in Kenya.

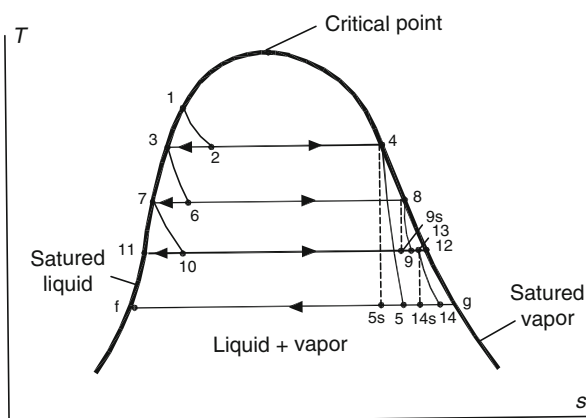
Two-Phase Geothermal Power Station In the majority of worldwide geothermal fields, the geothermal fluid is separated in an aboveground separator into a separate stream of brine and steam. In a low to moderate enthalpy resource, the steam quality is 10–30% of the entering fluid quality if comparing enthalpy and separation pressure. The two streams can efficiently be utilized in a two-phase geothermal station as shown in Fig. 87. Separated steam (usually with some percentage of noncondensable gases or NCGs) is introduced in the evaporator to vaporize the organic fluid [81].

The geothermal condensate is mixed with the separated brine to provide the preheating medium of the organic fluid. In the ideal case, presented in the T-Q diagram (Fig. 88), the latent steam heat equals the heat of vaporization of the organic fluid and the sensible heat of the brine plus condensate equals the heat





Geothermal Power Conversion Technology. Figure 75
Combined single- and double-flash station [35]

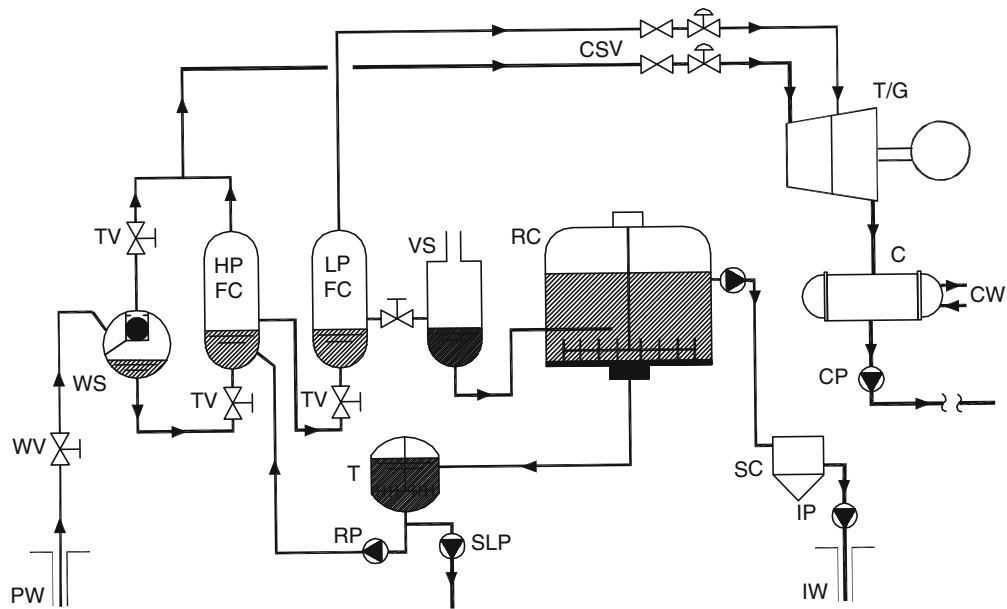


Geothermal Power Conversion Technology. Figure 76
Process diagram for combined single- and double-flash station

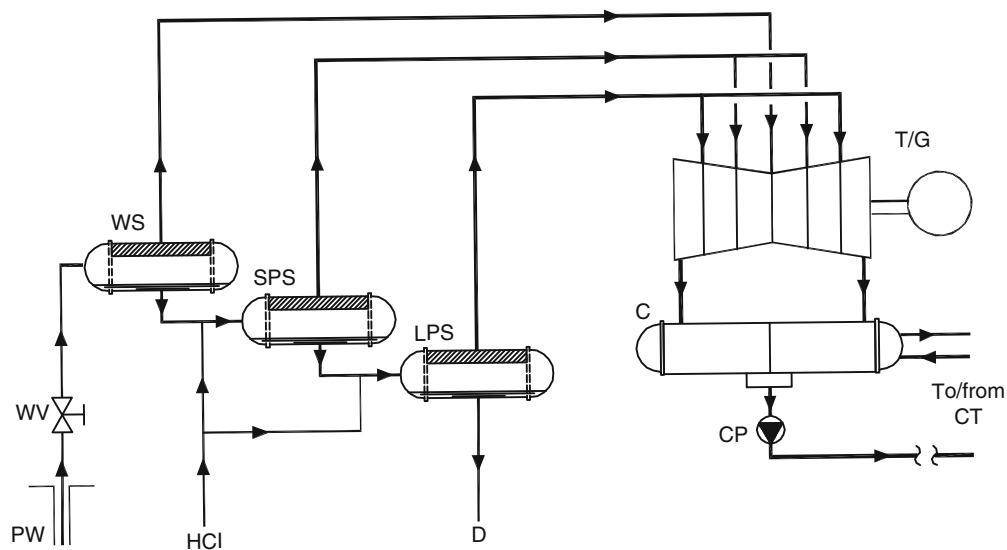
turbines and condensed at an intermediate pressure (and temperature).

The condensed vapor preheats the main organic fluid stream as it exits the recuperator. The extracted organic fluid forms a secondary cycle which generates an additional 5–8% electrical power. When there is extra steam compared to brine (higher enthalpy) the above cycle is effective and the cooling temperature of the brine plus condensate is limited.

Figure 90 is a flow temperature diagram of the higher enthalpy cases. Line A is the simple two-phase cycle preheating phase. The significant irreversibility is represented by the large space between the steam and brine lines and line A. Line B shows the preheating phase in a recuperated two-phase cycle. Here irreversibility is reduced and the cycle efficiency is increased accordingly.



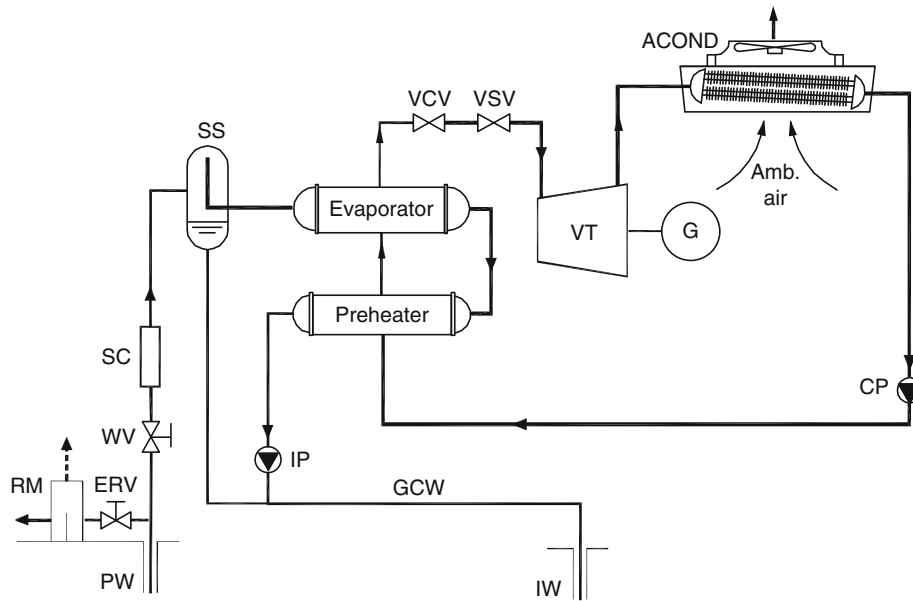
Geothermal Power Conversion Technology. Figure 77
schematic flow diagram for a FCRC power station, after [35, 69]



Geothermal Power Conversion Technology. Figure 78
Flow diagram for pH-Mod power station

The third line, C, demonstrates the additional gain in efficiency by using the two-phase/extraction cycle. The line moves further to the right, thus decreasing the gap between the heating line and the working fluid line.

Another indication of efficiency increase from cycle A to B and to C, is the increasing heat quantity for heating the working fluid, as presented by points QA, QB, and QC.



PW - Production Well

IW - Injection Well

RM - Rock Muffler

WV - Wellhead Valve

SS - Steam Separator

SC - Scrubber

VCV - Vapor Control Valve

VSV - Vapor Stop Valve

G - Generator

VT - Vapor Turbine

CP - Condensate Pump

IP - Injection pump

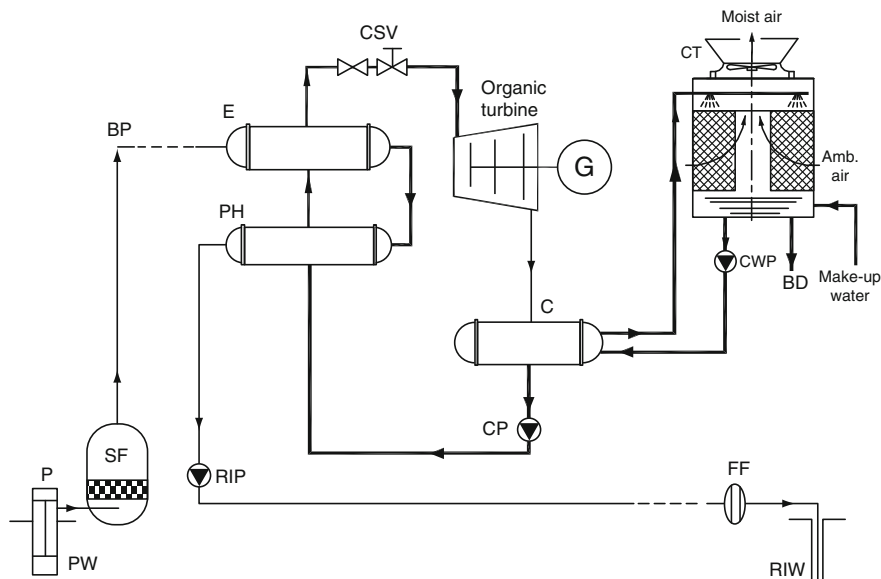
CW - Cooling Water

ACOND - Air Cooled Condenser

GCW - Geothermal Cooled Water

Geothermal Power Conversion Technology. Figure 79

Binary cycle operated by flashed steam



Geothermal Power Conversion Technology. Figure 80

Simplified schematic of a water-cooled binary geothermal power station [35]

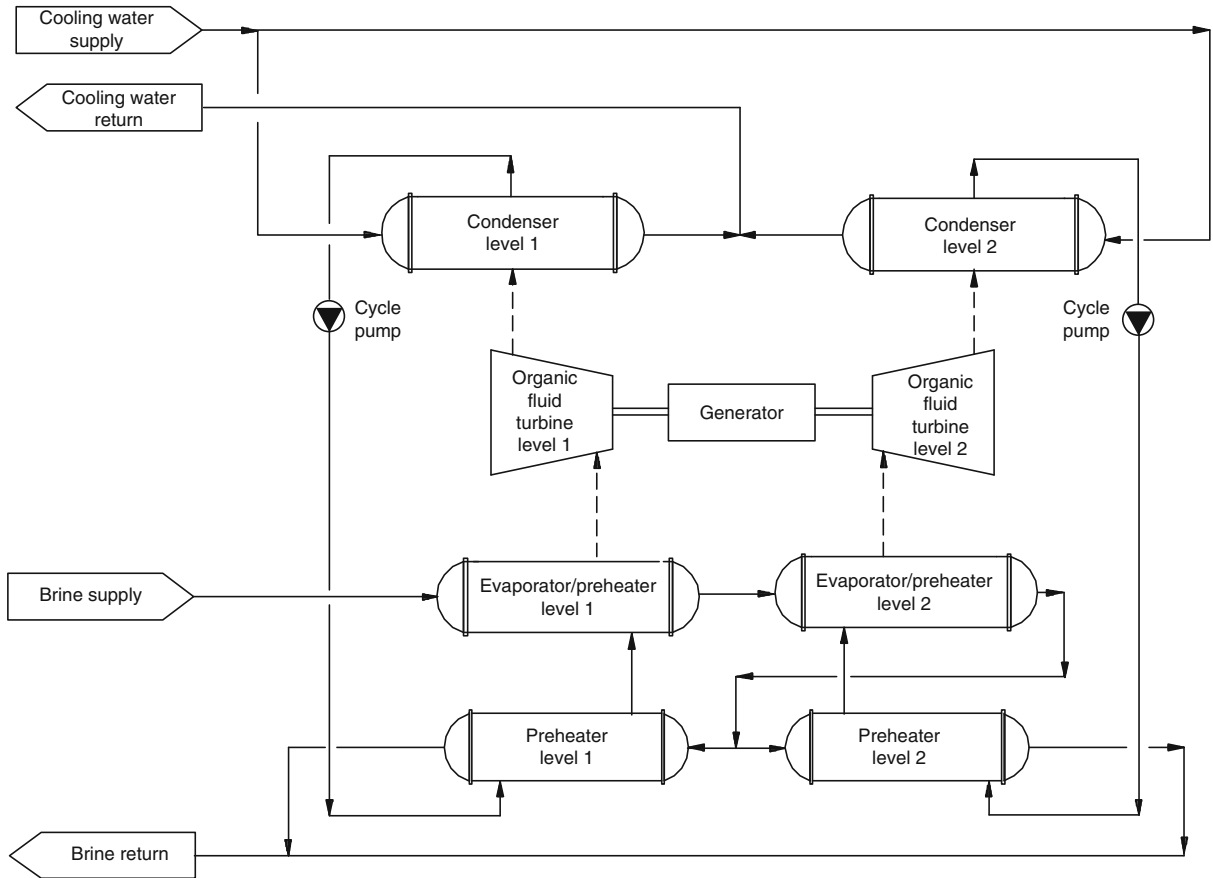


Hybrid Geothermal Cycles

Use of a back pressure steam turbine

For this case, assume that a single-flash station has been running for some time (usually a few years), and the reservoir has shown itself capable of sustaining operations for many more years. The power output can be raised by adding a binary unit between the separators and the reinjection wells. A simplified schematic of such an arrangement is given in [Fig. 91](#).

The thermodynamic process coordinates to facilitate comparison with the cycles in the previous sections. The power units are coupled thermodynamically through the preheater FH and the evaporator E. Using the state points of Fig. 91, the First Law gives the relationship between the brine flow rate \dot{m}_b , from the



Geothermal Power Conversion Technology. Figure 82
Integrated Two Level (ITLU) Power Station

wells (state 1) and that of the Organic Rankine cycle working fluid \dot{m}_{wf} :

$$\dot{m}_b(1 - x_2)c_b(T_3 - T_2) = \dot{m}_{wf}(h_a - h_c) \quad (82)$$

This equation shows that the heat extracted from the waste brine is equal to the heat absorbed by the Organic Rankine cycle working fluid, assuming perfect insulation on the heat exchangers. After solving for the working fluid flow rate it is found:

$$\dot{m}_{wf} = \dot{m}_b(1 - x_2) \left[\frac{c(T_3 - T_2)}{h_a - h_c} \right] \quad (83)$$

Since the state points 1, 2, and 3 for the flash unit are fixed and the new state point 7 is subject to the constraint imposed by silica precipitation, only the Organic Rankine cycle parameters are open for assignment.

Geothermal Combined Cycle Power Stations In case of high enthalpy dry steam or vapor dominated sources the use of condensing steam turbines present a number of disadvantages. First the high humidity in the many stages of the low pressure turbine portions lead to efficiency loss and erosion/corrosion of the blades. Secondly if non-condensable gases are present use of vacuum pumps is necessary to avoid efficiency loss due to back pressure and reduction of the heat transfer coefficient of condensation. Using only the high pressure part of the condensing steam turbine (also called “back pressure steam turbine”) and using the exhaust steam as the heat source for the evaporator of an Organic Rankine cycle [82], we get a geothermal combined cycle Fig. 94 which avoids the above drawbacks of condensing steam turbines.



Geothermal Power Conversion Technology. Figure 83
Photo of 40 MW Heber ITLU Power station (Courtesy of ORMAT)

The collected brine at the moisture remover MR exit is dumped (if small quantity), or added to the condensate exit or delivered to the reinjection wells, depending on the quantities and temperatures of the brine.

The T-Q diagram of the Organic Rankine cycle (Fig. 95) will be the same as given before for the two-phase power station.

Geothermal combined cycle configuration avoids both drawbacks: steam expansion in the back pressure steam turbine is smaller limiting the wetness of the steam and its effects while the partial pressure of the non-condensable gases (NCG) is small and so is its effect on the condensation in the condenser/vaporizer of the Organic Rankine portion of the cycle. An additional advantage is that the NCG is above the atmospheric pressure, therefore can be ejected without the need of vacuum pumps or reinjected with the condensate into the injection well. Another advantage is that the use of an air-cooled condenser on the ORC is more cost effective than on a condensing steam turbine.

Use of a Back-Pressure Steam Turbine Another approach for a moderate enthalpy two-phase heat source is the use of a back-pressure steam turbine

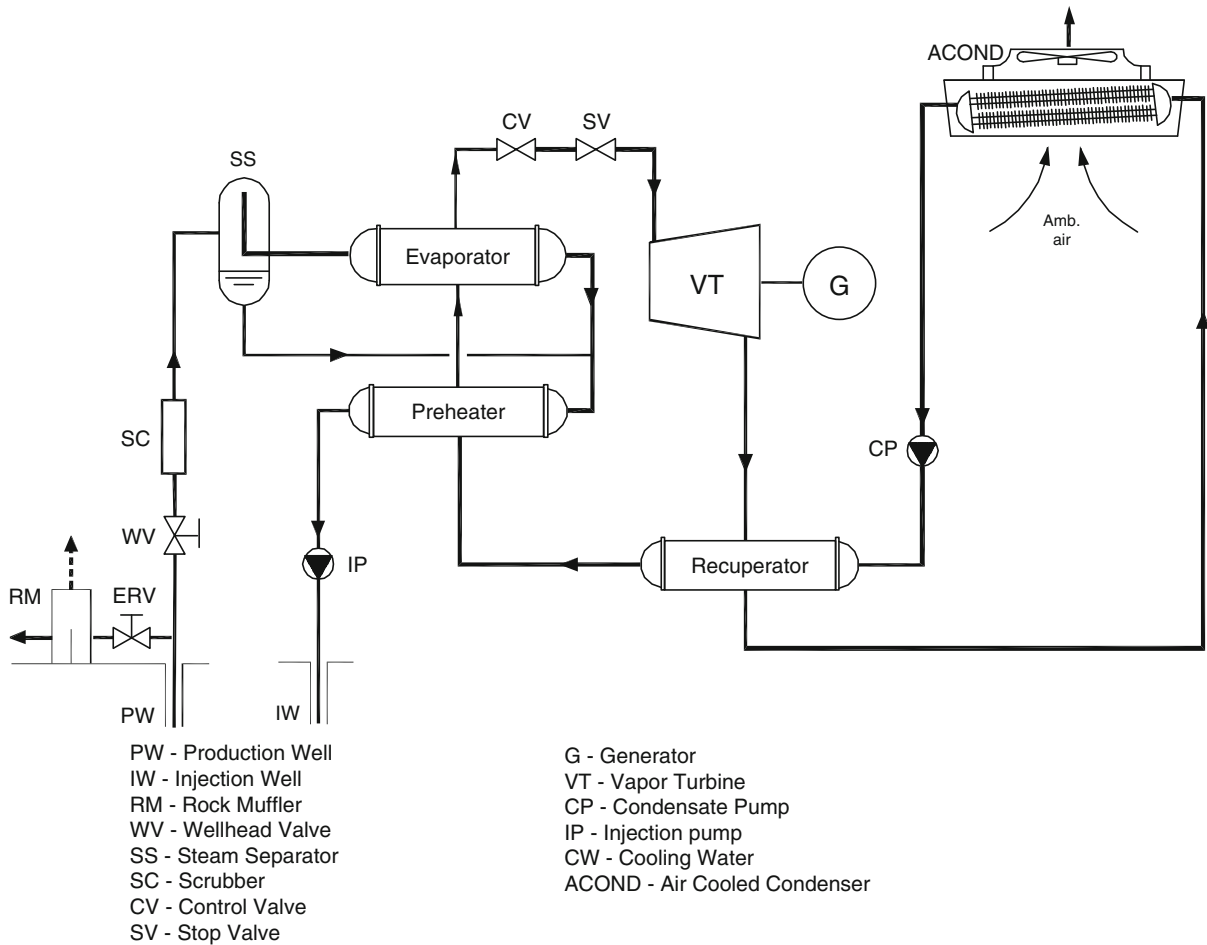
which generates extra power from excess steam not required for the ORC evaporator.

Low-pressure steam exiting the back-pressure steam turbine (Fig. 96) is used to partially preheat the organic fluid.

The gap between the organic fluid steam and the preheating line could be filled more efficiently by a multistage (two or more) back-pressure steam turbine, with steam extraction between the stages. The number of stages takes into account the process trade-off optimization between higher efficiency and the complication (and cost) of the system.

A system based on the above cycle is now operating in the 20 MW Amatitlan geothermal project in Guatemala (station photo is given in Fig. 97 and the station scheme in Fig. 98).

Geothermal Integrated Combined Cycle Power Stations When a bottoming Organic Rankine cycle is integrated with an air-cooled combined cycle, the result is a station with practically zero emissions. An integrated CC single-flash-binary station is shown schematically for water-cooled system in Fig. 99 [73] while Fig. 100 shows the same concept for air-cooled system [78] with three separated turbines. The process diagram is in two parts, the main upper combined cycle



Geothermal Power Conversion Technology. Figure 84
 Recuperated Organic Rankine cycle in two-phase binary power station

and the bottoming ORC portion. Geothermal steam first drives the back-pressure steam turbine and then is condensed in the upper ORC evaporator (E in Fig. 99). The two turbines in the upper part of the station may be connected to a common generator.

The separated brine (state 3) is used to preheat and evaporate the working fluid in the bottoming ORC. Noncondensable gases flow with the steam through the steam turbine ST into the evaporator where they are isolated, removed and compressed for recombination with the waste brine prior to reinjection. The brine holding tank (BHT) collects all the steam condensate, waste brine and compressed gases that go back into solution (Fig. 99).

In principle, this station has no emissions to the surroundings. The only environmental impact is the

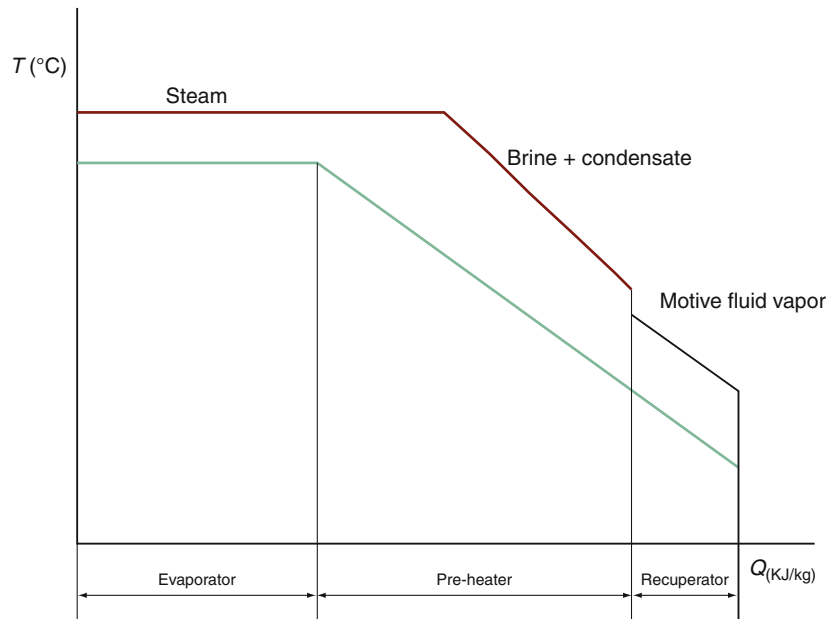
heat rejected to the atmosphere from the Organic Rankine cycle condensers. The scheme shows air-cooled condensers but water-cooling is an option.

This configuration was first used in 1992 in the 30 MW Puna power station in Hawaii, then in the 125 MW Upper Mahiao in the Philippines (Fig. 101), 100 MW Mokai 1 and 2 in New Zealand.

An example for an integrated combined cycle station is given also by DiPippo [73].

Combined Heat and Power

Iceland In most geothermal sites the option for utilization of the residue heat energy contained in the waste brine does not exist. The main reason is the distance from population centers. Iceland is one of



Geothermal Power Conversion Technology. Figure 85
Recuperated ORC in two-phase binary power station

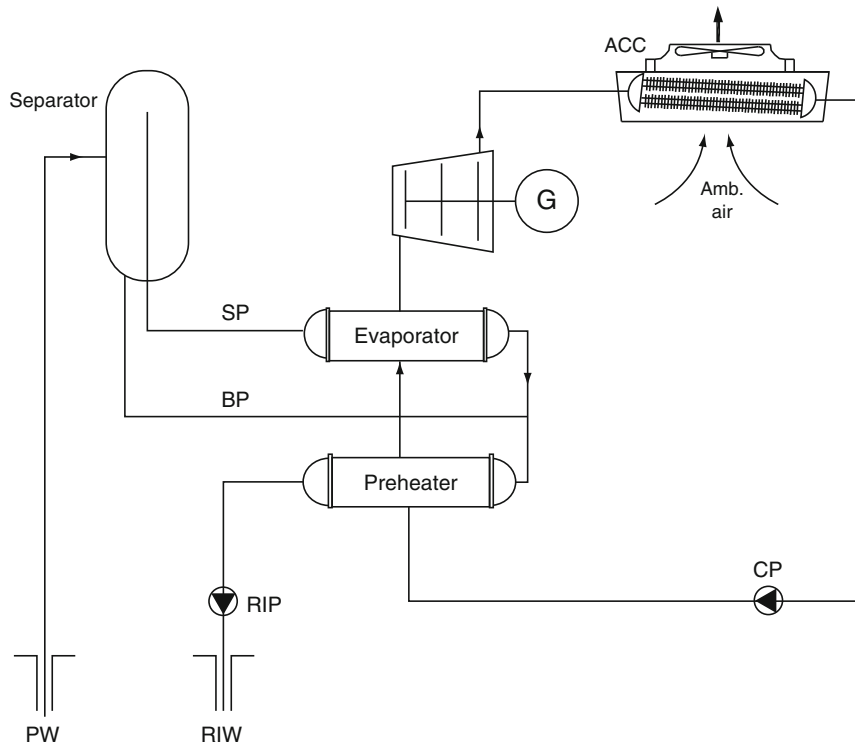


Geothermal Power Conversion Technology. Figure 86
Two-phase 14 MW Ribeira Grande power station in the Azores (Courtesy of ORMAT)

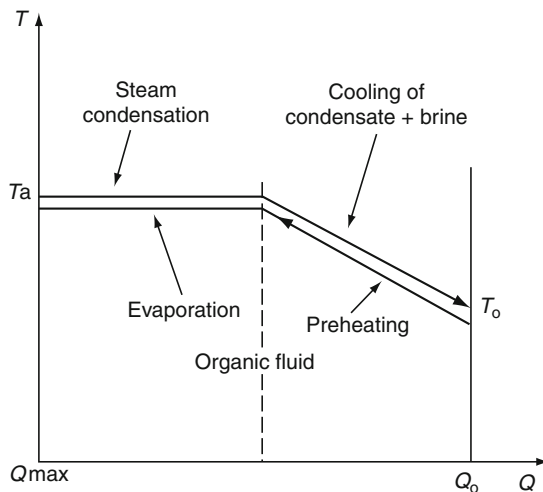
few examples where heat can be used for district heating and similar usage.

The Svartsengi geothermal area is close to the town of Grindavik on the Rekjanes peninsula and is part of an

active fissure swarm, lined with crater-rows and open fissures and faults. The high-temperature has an area of 2 km^2 and shows only limited signs of geothermal activity at the surface. The reservoir contains much



Geothermal Power Conversion Technology. Figure 87
Two-phase binary power station



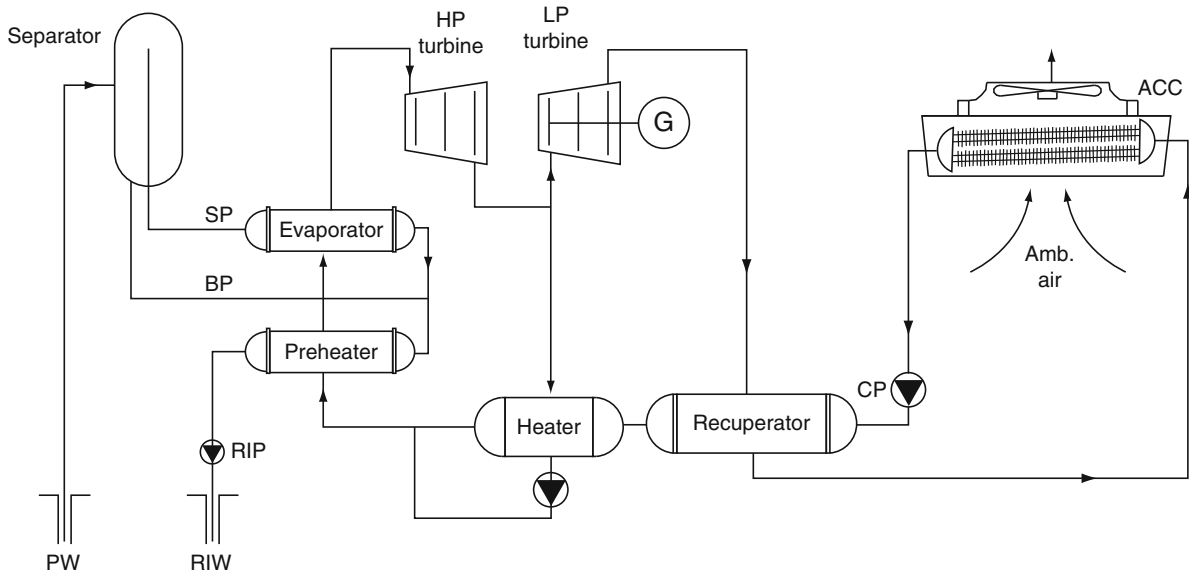
Geothermal Power Conversion Technology. Figure 88
T-Q diagram of a two-phase binary power station

energy with at least 8 wells supplying the Svartsengi Power Stations with steam [79]. The steam is not useable for domestic heating purposes and heat

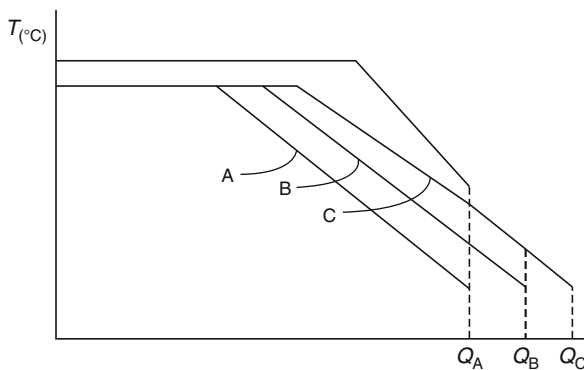
exchangers are used to heat cold groundwater with the steam. Some steam is also used for producing 16.4 MW_e of electrical power, see Fig. 102 bottom. Figure 102 top shows the distribution system piping of hot water to nine towns and the Keflavik International Airport. The effluent brine from the Svartsengi Stations is disposed of into a surface pond, called the Blue Lagoon. This is popular for tourists and people suffering from psoriasis and other forms of eczema seeking therapeutic effects from the silica-rich brine.

In 1969, the Grindavik municipal council decided to do a study of harnessing geothermal energy in the Svartsengi area to heat houses in the village. The wells drilled at that time, 240 and 430 m deep, looked very promising. There was some disappointment as it was revealed that:

- This was a high-temperature geothermal area (i.e., with temperatures rising to more than 200°C at less than 1,000 m depth (to hot for domestic usage).
- The geothermal reservoir contained water with about two thirds of the salinity of the sea.



Geothermal Power Conversion Technology. Figure 89
Secondary Organic cycle with LP partial vapor admission



Geothermal Power Conversion Technology. Figure 90
T-Q diagram of the high enthalpy secondary organic cycle

Due to the level of salinity and the high temperature of the water, it was clear that it would not be possible to utilize the geothermal fluid directly as had been the case in Reykjavik and most other places in Iceland. What was needed was the development of a method of heat exchange to facilitate the utilization of the geothermal power.

Rogner Hotel in Austria The 250 kW geothermal project at Bad Blumau is the first geothermal project developed in Austria by the private sector following the

deregulation of the electricity industry in this country. Besides its private ownership structure, the project is unique due to its ability to generate electrical power and district heating by using a low-temperature geothermal resource. The unit is shown in Fig. 103.

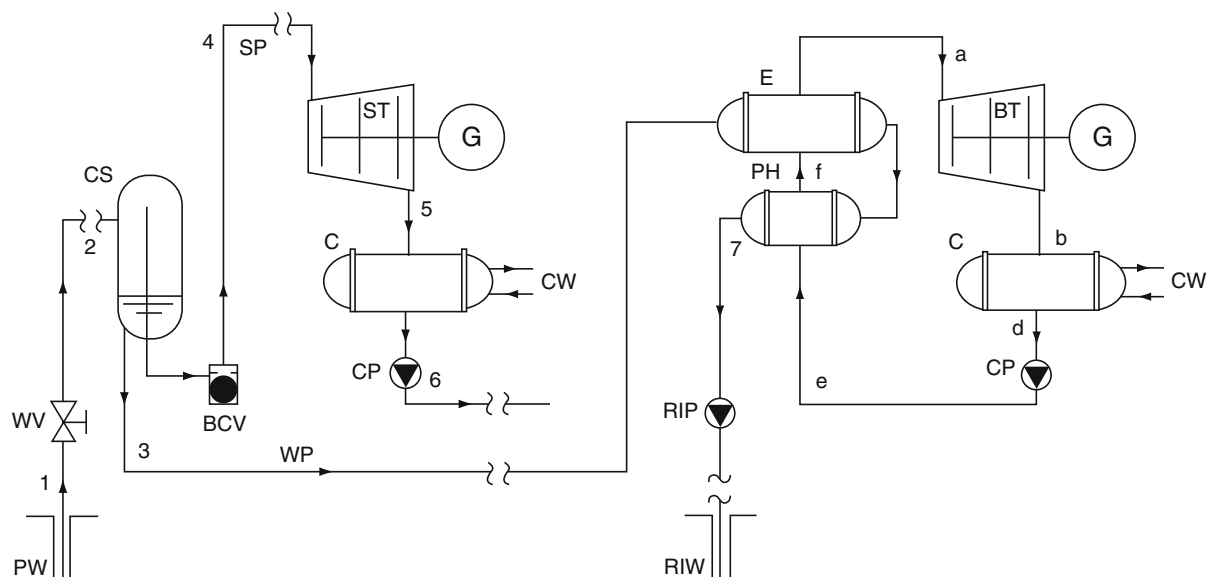
The air-cooled ORMAT® Energy Converter (OEC) CHP module has been in commercial operation since July 2001. With an annual availability exceeding 99%, the station delivers about 1,300,000 kWh annually to the local grid. The geothermal CHP module utilizes brine at $\sim 110^\circ\text{C}$ available from a 3,000 m deep production well. Exiting the OEC unit at a temperature of $\sim 85^\circ\text{C}$, the brine is then fed into the district heating system providing heat for the Rogner Bad-Blumau Hotel and Spa. The geothermal brine is returned from the district heating system and injected into a 3,000 m depth well.

The system is a pollution-free, unattended operating power generation module, which averts about 1,000 tons of CO_2 emissions annually.

Experimental Power Stations

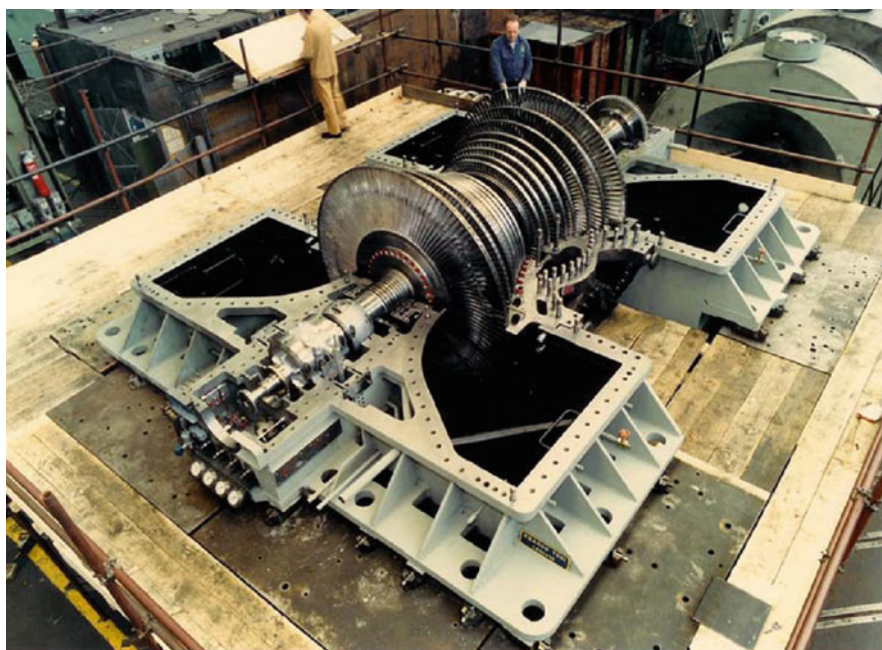
Dual-Fluid Organic Rankine Cycle

The first binary station in the USA was the Magmamax station at East Mesa in California's Imperial Valley.



Geothermal Power Conversion Technology. Figure 91

Combined single-flash and ORC station; after [69]



Geothermal Power Conversion Technology. Figure 92

Momotombo Franco Tosi 35 MW steam turbine (Courtesy of Franco Tosi)

The station was a 12.5 MW station that began operation in 1979 using a dual-fluid cycle (two different hydrocarbons were used in interlocking Rankine cycles).

One a subcritical cycle and the other a supercritical cycle [83, 84]. The typical dual fluid system is shown in Fig. 104.



Geothermal Power Conversion Technology. Figure 93
Ormat 9 MW ORC bottoming power unit (Courtesy of ORMAT)

As with the dual-pressure cycle, incentive here is to create a good “match” between the brine and the working fluid heating-boiling curves. The temperature-heat transfer diagram [Fig. 105](#), shows this relationship. The discontinuity between state points 5 and 11 arises from the internal heat transfer between the working fluids and does not involve the brine. From the diagram it is seen that the pinch point occurs between state b on the brine cooling curve and state 6, the bubble point for fluid 1. The near-parallelism between the brine and the working fluids in the preheaters means that the thermodynamic irreversibilities will be low, as will the loss of energy during the heat transfer process in those components. Since the average temperature difference in the fluid 1 evaporator is relatively large, it will be associated with a higher energy loss.

If fluid 1 is raised to a supercritical pressure before entering its preheater, the temperature-heat transfer diagram would change dramatically, see [Fig. 106](#).

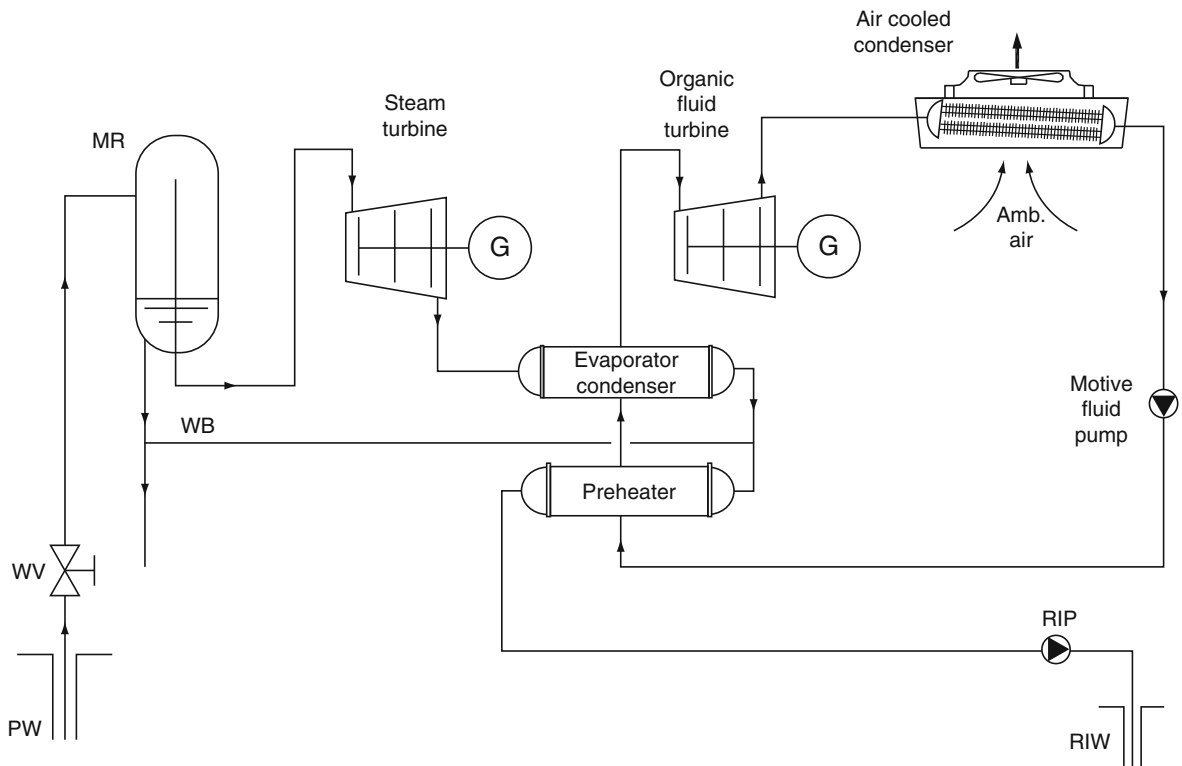
The sharp corner at state 6 denoting the bubble point for fluid 1 has vanished. Fluid 1 now has a smooth heating curve taking fluid from a cool compressed liquid to a hot supercritical vapor. There will

still be a point of closest approach between the two curves, but it is far less pronounced. This allows a good match between the brine and the working fluids which results in lower energy losses and higher utilization efficiency for the cycle.

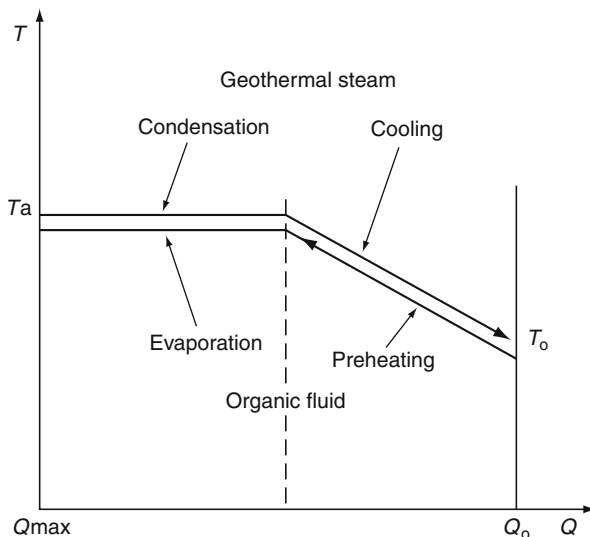
As already mentioned, the supercritical cycle has higher thermal efficiency. However, the pump work is using greater fraction of the net cycle work and is about 50% higher than for the subcritical cycle.

There are additional practical difficulties with a supercritical cycle. The higher pressures may require change of traditional use of shell and tube heat exchangers in the geothermal application where the brine flows in the tubes and the organic fluid in the shell side. This allows for practical operation of in-tube cleaning as may be required in brine flow during long operation. Also, once it is changed, thicker and more costly tubing in the heat exchangers is required.

In both cases, the heat recovered from the condensation of fluid 1 is used for evaporation of the second fluid in E2. In a T-Q diagram between the two fluids there will be two parallel lines as given in [Figs. 105b](#) and [106b](#).



Geothermal Power Conversion Technology. Figure 94
Geothermal Combined cycle (GCC)



Geothermal Power Conversion Technology. Figure 95
T-Q diagram of the ORC part of the CC

This was one reason why the original Magmamax station [83] placed the supercritical isobutane inside the tubes and the brine on the shell side of the heat exchangers.

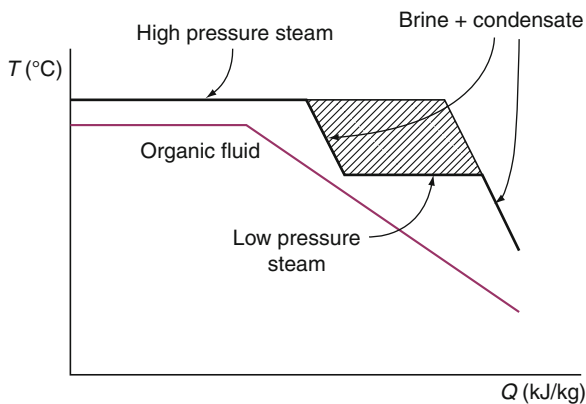
Kalina Cycle

Water-ammonia mixtures have long been used in absorption refrigeration cycles [85]. It was not until Kalina patented his Kalina cycle [86] that this working fluid was used for power generation cycles. A typical Kalina cycle, KCS-12, is shown schematically in Fig. 107. The features that distinguish the Kalina cycles (there are several versions) from other Organic Rankine cycles are as follows:

- The working fluid is a binary mixture of H_2O and NH_3 .
- Evaporation and condensation occur at variable temperature (requires several heat exchangers).

- Cycle incorporates heat recuperation from turbine exhaust.
- Composition of the mixture may be varied during cycle in some versions.

As a consequence, Kalina cycles show improved thermodynamic performance of heat exchangers by reducing the irreversibilities associated with heat transfer across a finite temperature difference. The heaters are arranged so a better match is maintained between



Geothermal Power Conversion Technology. Figure 96
Preheating using exhaust in a back-pressure steam turbine

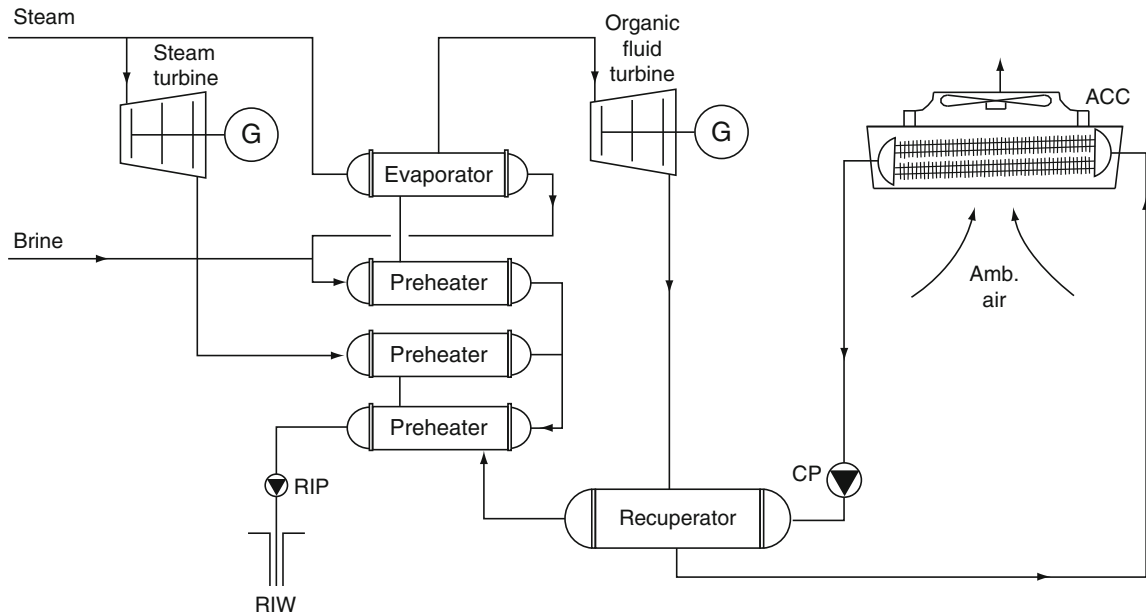
the brine and the mixture at the cold end of the heat transfer process (where improvements in energy preservation are most valuable).

A reheater is needed because the water-ammonia mixture has a normal saturated vapor line, i.e., $dT/ds < 0$, leading to wet mixtures in the turbine. The station relies on good heat exchangers because more heat is transferred than in a supercritical binary station of the same power output. Bliem and Mines [87] showed that the Kalina cycle of Fig. 107 requires about 25% more heat transfer. A possible advantage to using the recuperative preheaters is that they reduce the heat load on the condenser and cooling tower. The lower capital cost of a smaller condenser and cooling tower must be compared to the extra cost for the recuperators. Over the long haul, the resulting higher efficiency should mean lower operating costs.

The station is more complex than a basic binary station, particularly when a distillation column is used to vary the mixture composition. The simplest configuration of the Kalina cycle with variable working fluid composition is shown in Fig. 108. The separator S allows a saturated vapor rich in ammonia to flow to the turbine, thus permitting a smaller and less costly turbine than for a hydrocarbon working fluid.



Geothermal Power Conversion Technology. Figure 97
20 MW Amatitlan Power Station in Guatemala (Courtesy of ORMAT)



Geothermal Power Conversion Technology. Figure 98
Block diagram of the Amatitlan Power Station

The weak solution (a liquid rich in water), is used in the preheater and is then throttled down to the turbine exhaust pressure before mixing with the strong solution to restore the primary composition. The mixture is then used in a recuperative preheater RPH prior to being fully condensed.

A possible difficulty for the Kalina cycle striving for high efficiency, is maintaining very tight pinch-point temperature differences in the heat exchangers. Also, the advantage of variable-temperature condensation is lessened because the condensing isobars of the ammonia-rich $\text{NH}_3\text{-H}_2\text{O}$ mixtures used in power cycles concave upward, leading to a pinch-point. Thus, there are relatively large temperature differences at the beginning and end of the condensing process.

DePippo compared the Kalina cycle with a simple ORC cycle [88] (Second Law comparison) and concluded that for low temperature brine the Kalina cycle is about 26% more efficient than the ORC cycle. Paola Bombarda [89] compared the Kalina cycle against ORC and found that the dominant factor bringing high efficiency to the Kalina cycle is the system operating pressure.

Kalina cycles operating at pressures lower than 100 bar will have lower efficiency than the ORC system,

while those above 100 bar have higher efficiency. This is a disadvantage to the Kalina cycle as high pressure systems are likely to be more expensive in addition to the handling of >100 bar ammonia mixtures.

Geopressured Geothermal Systems

There are deep reservoirs holding geofluids at high pressure and temperature. Those of interest have pressures about 200–300 bar and temperatures ranging between 110°C and 200°C . The water, located at a depth of 3–5 km is high-salinity brine with large amounts of methane dissolved in it.

One location is the Gulf of Mexico, which is known because of the extensive oil/drilling that occasionally ends with water-dominated fluid instead of oil. The high pressurized liquid can drive a hydraulic turbine, then flow through binary unit evaporator and preheater to produce additional power. The solubility of methane depends on pressure, temperature and the salinity of the brine. Extracting the methane after the hydraulic turbine and transferring most of its heat to the organic liquid may give the best results. Test drills along the Louisiana coastline [90] and one actual methane extraction test made by the DOE in Pleasant Bayou



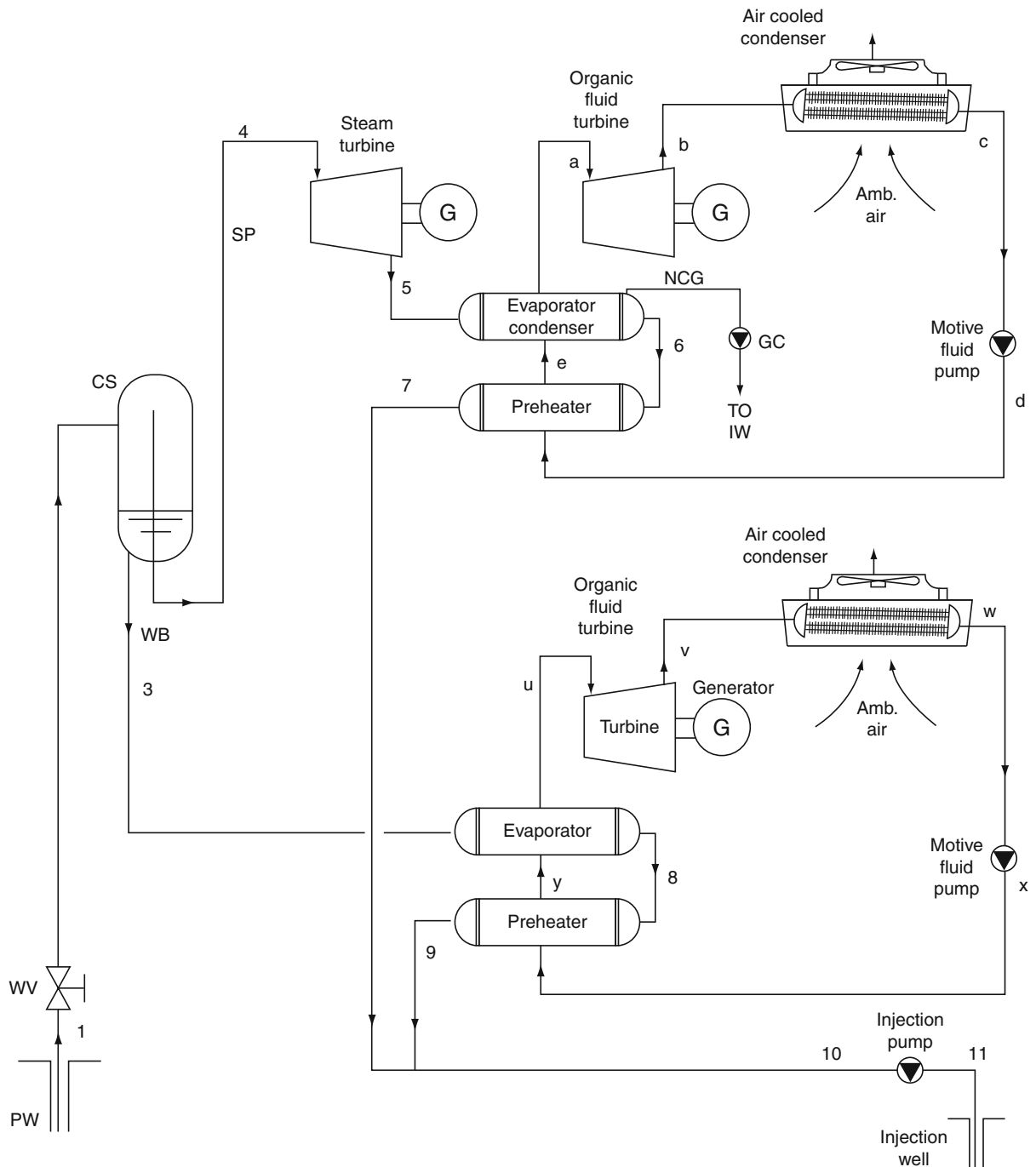
G

G

G

G

G



Geothermal Power Conversion Technology. Figure 100

Integrated geothermal combined cycle and bottoming ORC Power Station – air-cooled system

Hybrid Geothermal Fossil Fuel Power Station

Geothermal steam, whether dry or flashed, is bound to expand into the wet zone and therefore its power

production is somewhat limited. The idea of using fossil fuels for superheating and to enhance geothermal resources is not a new concept. A paper published in



Geothermal Power Conversion Technology. Figure 101

125 MW Upper Mahiao Geothermal Power Station in the Philippines (Courtesy of ORMAT)

1924 indicated that it was already suggested by P. Caufourier [95]. He proposed a hybrid power system in which hot water from a geothermal spring would be successively flashed four times and the generated steam superheated in a fossil fired superheaters prior to being admitted to a multi-pressure turbine. Thermodynamic analysis by DePippo [96] showed that the system which burns fuel for that purpose only is not economical. A different approach suggested usage of heat recovery of a gas-operated gas turbine for superheating of geothermal steam in The Geysers geothermal field [97]. Such a hybrid system would have the highest utilization of fossil fuel as compared with regular GT power generation. For additional information on superheating of geothermal steam see studies by Brown University [98, 99].

In geopressured wells there is usually some natural gas content that is separated before or after usage of the heat energy in binary or flash steam cycles. The gas can be used in a gas engine such as gas turbine and the exhaust heat utilized for the geothermal steam superheating as previously mentioned. Such systems were analyzed by Chang and Williams [100] in a work sponsored by the DOE 1985.

Hybrid Geothermal Biomass Power Station

The concept is similar to the hybridization with fossil fuel, but because of generally lower combustion temperature of biomass, the thermodynamic draw-back is

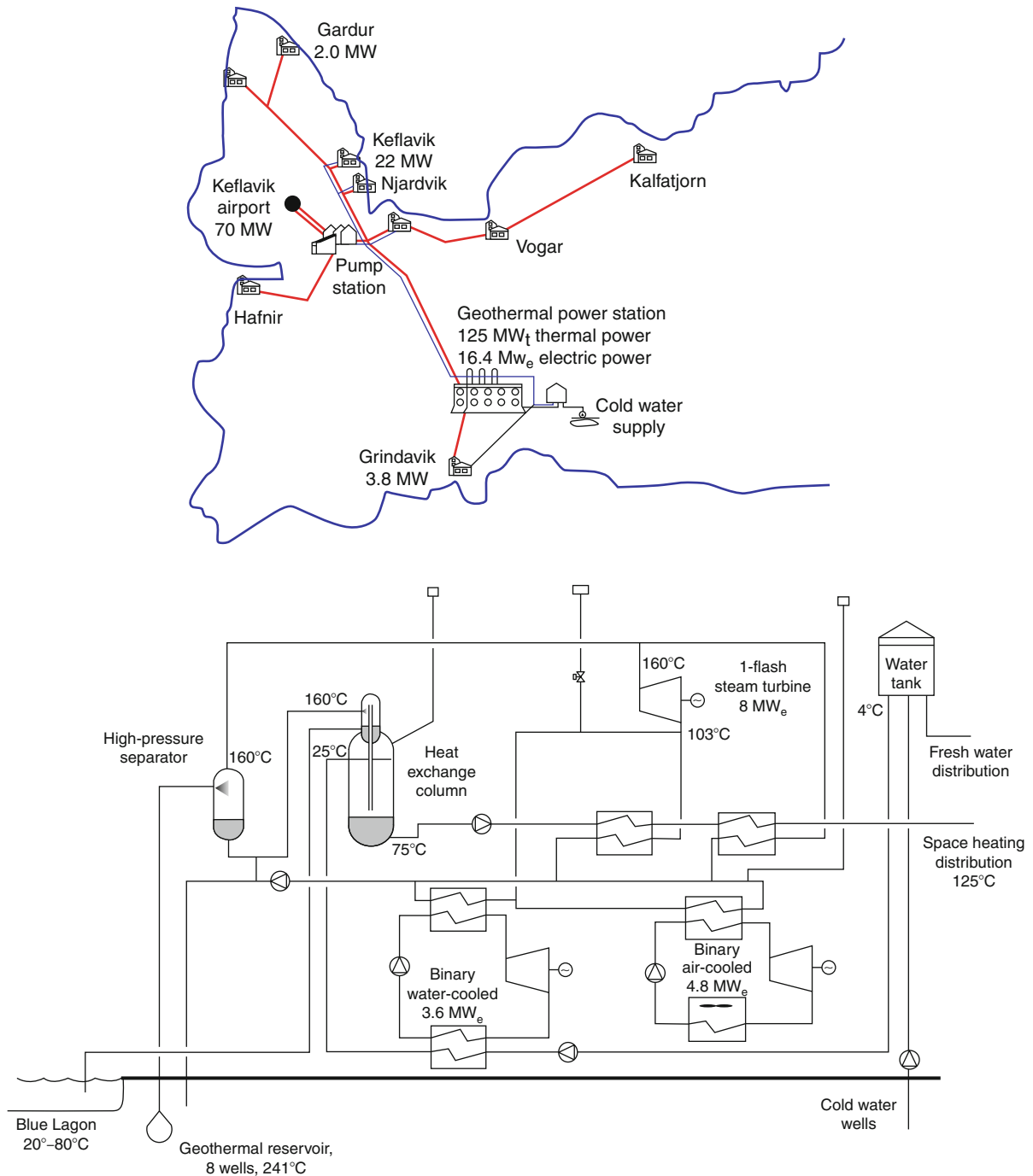
smaller. The geothermal resource is used to preheat the motive fluid while the biomass is used to evaporate the motive fluid. Although the exhaust heat from biomass combustion may provide all the preheating of the motive fluid making the geothermal resource use redundant, however because the high dew point of exhaust gases from biomass combustion there is room for the geothermal heat for preheating. A few such power stations were proposed and it seems that at least one was constructed.

Hybrid Geothermal Solar Power Stations

This hybridization presents three advantages. The first is thermodynamic, by providing all the heat of evaporation from the solar collector, the efficiency of the utilization of the sensible heat of the geothermal reservoir, is improved: see paragraph "Available energy" and Figs. 6 and 7. The second is an improvement of the load following of the station: more power produced during peak hours. Thirdly the economics is improved by a better use of the interconnecting facility and of the personnel. The outline of such a system is shown in Fig. 110.

Power Stations for Enhanced Geothermal Systems (EGS)

MIT report states that the majority of geothermal energy within drilling reach that can be utilized for



Geothermal Power Conversion Technology. Figure 102

The Sudurnes Regional heating system layout and flow diagram for Svartsengi Power Station

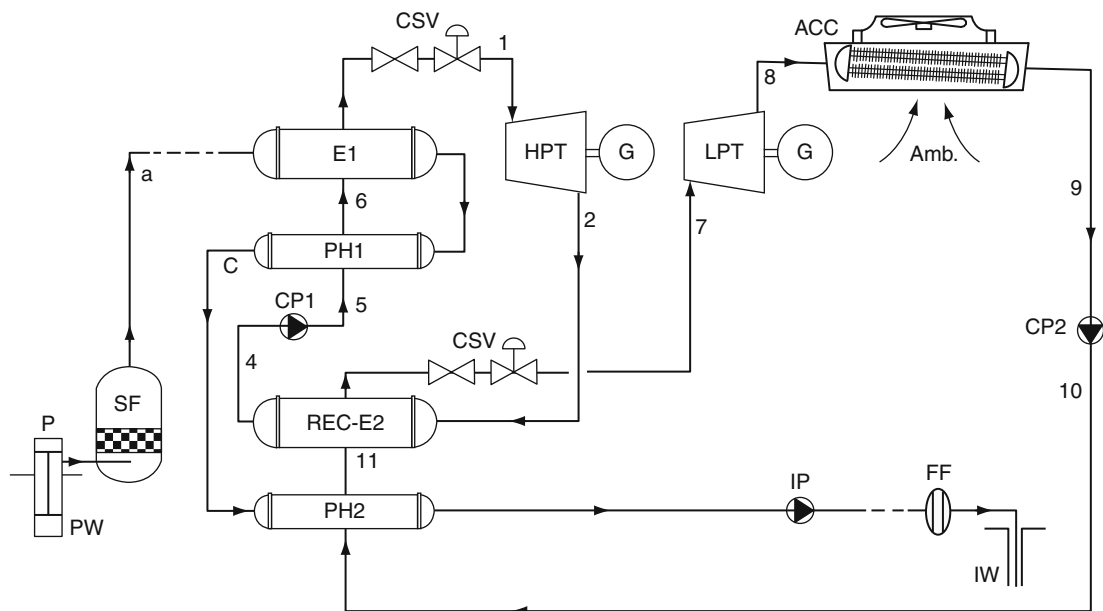
power production is in dry and nonporous rock [101]. Drilling into hot rock formations and creating cavities to accommodate large enough heat transfer area for

heating of water are considered as enhanced geothermal systems (EGS). Substantial progress has been made in developing and demonstrating certain components



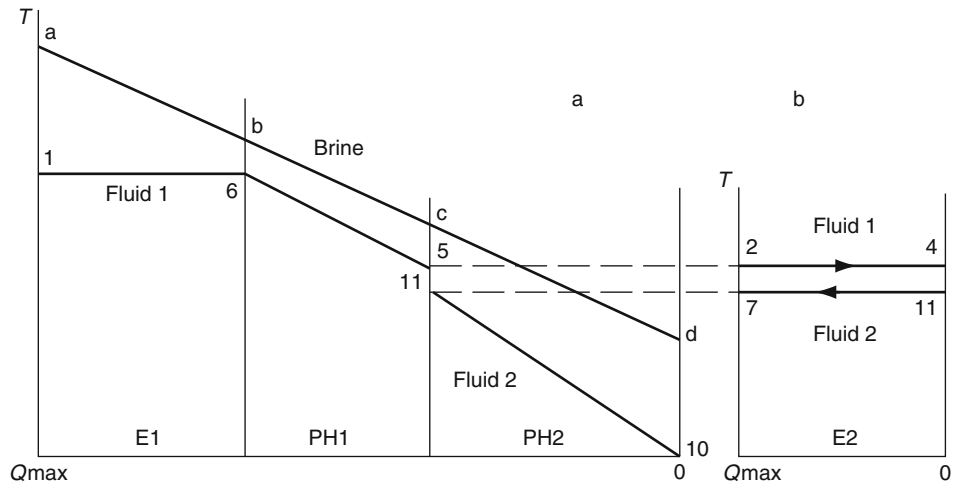
Geothermal Power Conversion Technology. Figure 103

250 kW Geothermal ORC Power Unit at Rogner Hotel and Spa, Bad Blumau, Austria (Courtesy of ORMAT)

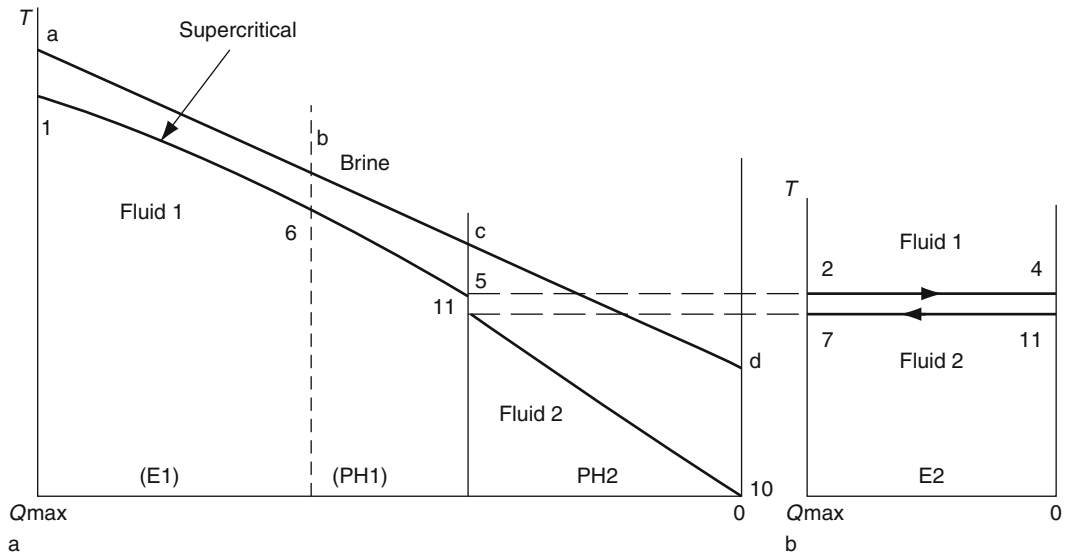


Geothermal Power Conversion Technology. Figure 104

Scheme of dual fluids binary power station



Geothermal Power Conversion Technology. Figure 105
T-Q diagram of dual organic fluids in subcritical condition

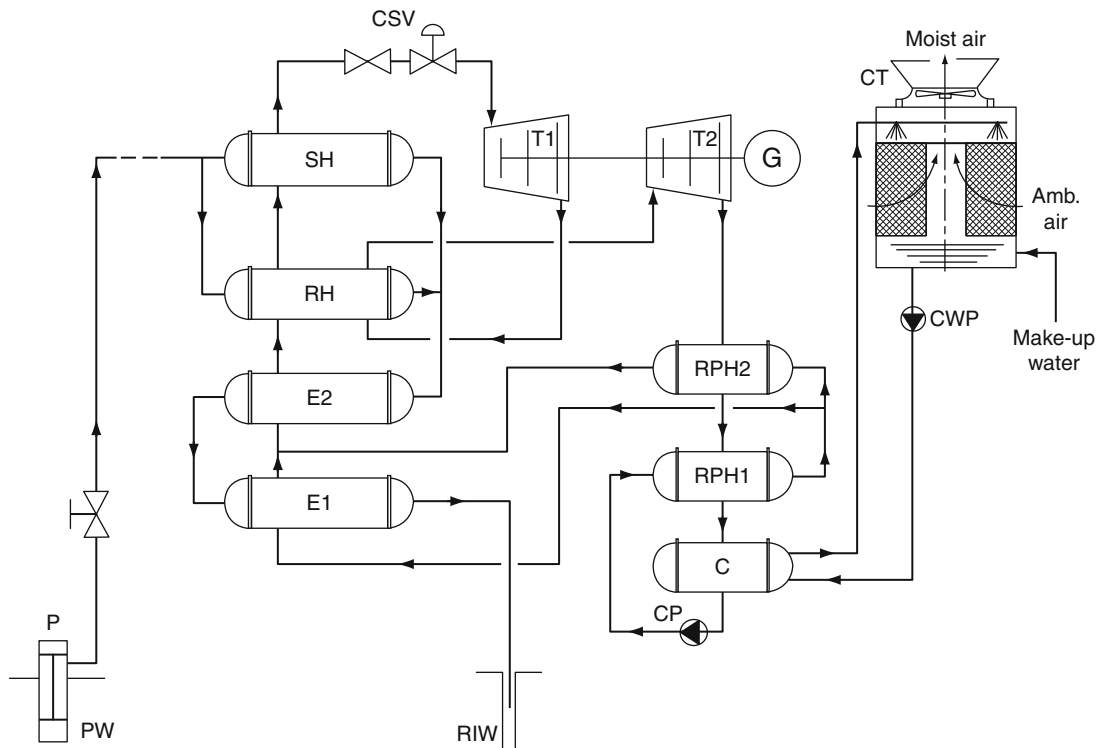


Geothermal Power Conversion Technology. Figure 106
T-Q diagram of dual organic fluids in supercritical condition

of EGS technology in the USA, Europe, Australia and Japan. Further work is needed to establish the commercial viability of EGS for electrical power generation, cogeneration and direct heat supply.

A separate, specific part of the present publication deals with the resource development, therefore attention is given here to the energy conversion system only.

Assuming that the build-up of the resource has been made. Water travelling through the fractures in the rock captures the rock heat and emerges from the production well accompanied by dissolved solids, large amount of particles and possibly noncondensable gases (NCG). Tests made in Japan [102–105], USA [106], UK [107] and other countries



Geothermal Power Conversion Technology. Figure 107

Typical Kalina cycle employing a reheater and two recuperative preheaters

range the out-coming water temperatures between 150°C and 270°C.

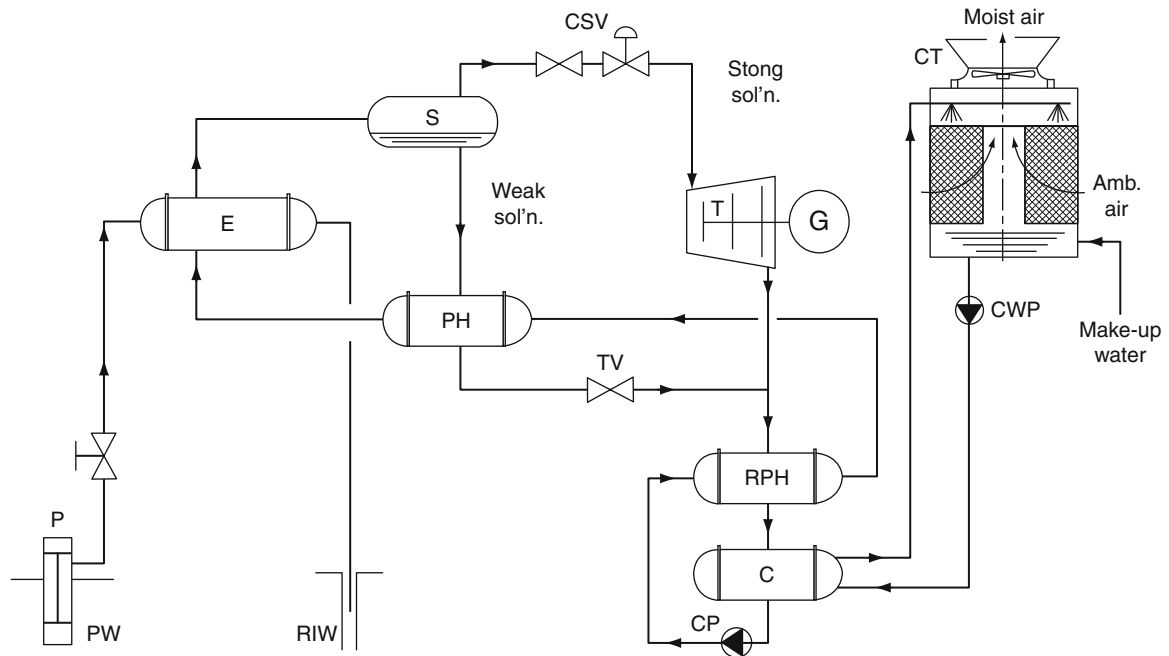
The lower-temperature brines are suitable for power generation via binary systems. The higher-temperature brines (above 200°C) are flashed, cured of NCG and particles and then can be directly used in steam turbines providing vapors free of aggressive components. However, due to the NCG problem and since the nature of dissolved materials in the brine may change in time, an indirect utilization by use of ORC binary stations is preferred. The air-cooled ORC stations are particularly well adapted to the EGSs. The somewhat higher installed cost of these systems is justified by environmental and long-term resource management considerations.

Most of the reported EGS stations are experimental with a status between planning, fundraising, drilling and partial operation (see list in [102]). The only partial EGS station in continuous commercial operation is in Landau, Germany using the Organic Rankine cycle [108].

Other Organic Rankine cycle stations are Soultz (France) [109], stations in Australia, Desert Peak and Newberry Oregon in the USA (being planned). The air-cooled ORC stations are particularly well adapted to the HDR/EGS. The fairly higher installed cost of these systems is justified by environmental and long term resource management consideration.

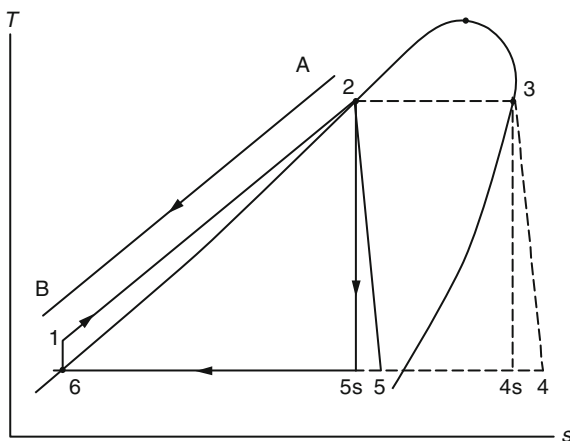
Trilateral Flash Cycle (TFC)

The Trilateral Flash Cycle (TFC) and its more recent “Smith” cycle [110] were developed for efficient utilization of HDR geothermal heat source. Because the work producing process is based on flash expansion of the liquid and the cycle is close to the thermodynamic trilateral ideal, it is a Trilateral Flash Cycle (TFC) system. The main feature as seen in Fig. 109 is to transfer the HDR heat (points A–B) to Organic liquid (points 1–2) and allow expansion in a two-phase expander (points 2–5) instead of the regular evaporator (points 2–3) and dry turbine



Geothermal Power Conversion Technology. Figure 108

Kalina cycle with variable composition of the water ammonia working fluid



Geothermal Power Conversion Technology. Figure 109
Scheme of the trilateral flash cycle (TFC)

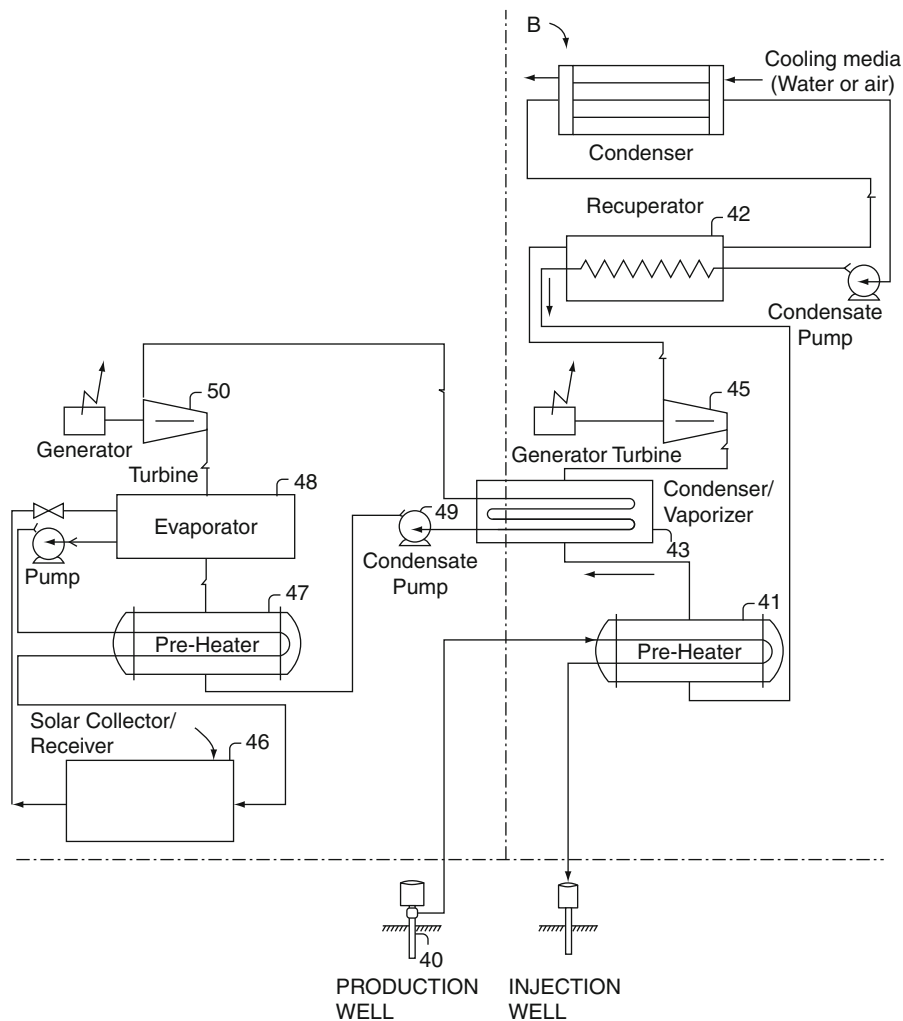
expansion (points 3–4). This is a drawback due to the relative low efficiency of the two-phase expansion resulting from operation deep in the wet zone. The more advanced Smith cycle suggests various options

of expansion procedures as the required ratio of expansion is above 100 and this cannot be achieved with the screw expander or even a radial turbine suggested by Smith. There is a two-phase expander in the first stage followed by separation vessel from which the system proceeds in two parallel lines, one a two-phase expander and the other a vapor turbine which expands in the dry zone. Theoretical analyses and cost estimates are optimistic, but the development of a large and still not constructed expander is still required. Few of the expanding devices suggested for this and similar “total-flow” cycles will be described in section on “[Total Flow Systems](#).”

Total-Flow Systems

Utilization of the steam or brine geofluids is accompanied by irreversibilities that consume a significant portion of the available energy and include quite large cost of equipment for separation, flashing and expansion. New ideas for direct use of the geofluid are basically directed at the large saving in equipment cost, direct expansion with half of turbine

A. TOPPING CYCLE: STEAM RANKINE CYCLE

B. BOTTOMING CYCLE:
ORGANIC RANKINE CYCLE

Geothermal Power Conversion Technology. Figure 110
Hybrid Geothermal Solar Power Station (Courtesy of ORMAT)

efficiency but no flashing. The main present directions are:

1. Single-stage impulse turbine [111]
2. Positive-displacement expander (a helical screw expander called also “bi-phase expander” and rotary expander) [112, 113]
3. Rotary separator turbine [114].

Two-Phase Turbine A hybrid geothermal power station comprises in addition to the geothermal heat source

a supplementary heat source such as biomass solar or even fuel. In a total-flow turbine the fluid expands from near the liquid line, while with a regular steam turbine the steam expands from near the saturation line. Comparing these two turbine cases between the same given temperatures (fluid high temperature and condensing temperature) it is found [111, 136] that the total-flow turbine efficiency can be about half the steam turbine efficiency for the same power production. Therefore steam turbines efficiency is in the range of 80%, with the best efficiency of single-stage impulse turbine still

at 23%. However, the single stage impulse turbine is remarkably smaller and the system uses all the brine energy compared with only the flashed steam energy.

Positive Displacement: Helical Screw Expander The helical screw machine is usually used as a compressor in refrigeration cycles. It is also used as air compressor in stationary and mobile compression units. Due to the screw shape of the twin rotors, pressure losses are small and high pressure ratio is obtained. If a high-pressure fluid is passed to the “exit” side of such a compressor then the expanding fluid will rotate the screw rotors. Tests with two-phase flow indicated that there are limitations to the rate of expansion caused mainly by the initial liquid content. Pre-flashing and partial flashing that changes the liquid–vapor ratio may diminish this issue. Tests conducted in the United States, New Zealand [112, 113] and in 1986 by Sprankle [115] were not continued.

The use of rotary expander used in passenger cars was considered by NEDO in Japan for the same task. A 300 kW prototype was built and tested in 1982, but there is no information on further tests.

Rotary Separator Turbine Here the two-phase flow (steam and brine) is separated in brine-driven primary turbine wheel. The steam is passed to the regular steam turbine for further expansion while the brine is drained from the system. The assumption is that the steam pressure is not significantly reduced by the above separation. The overall energy achieved from the two turbine wheels did not exceed the regular steam turbine. Tests performed in Roosevelt Hot Springs in 1981–1982 by Cerini et al. [114] were unsatisfactory. Additional work was summarized by Hughes [116] in 1986 with no reports in the following years.

Future of Geothermal Energy

Background

The cited papers used in this review do not always agree on the direction of development and forecast of power generation capacity but it is important to bring up and present all views to the reader.

United States

At the Symposium on energy sources for the future, in Oak Ridge, Tennessee, held between July 7–25, 1975,

M. King Hubbert [117] presented a Survey of World Energy Resources. He was very skeptical about the future of geothermal energy. The total world installed geothermal power capacity in 1975 was approximately 1,500 MW, and in spite of optimistic forecasts by geothermal power enthusiasts, such as Tester and Milora [118] at the same symposium, he forecasted an increase in only “order of magnitude.” As injection was hardly used at that time he claimed that in all likelihood most large installations will be comparatively short-lived, perhaps a century or so.

Today however, with deep-well explorations and successful injection programs there is a reassessment of the geothermal potential as reported at the World Geothermal Congress in Bali by Ruggero Bertani [123] in April 2010.

The report indicates that in the USA there are 9 states, all in the West, with operating geothermal power stations, and a total installed capacity of 3,093 MWe. The report estimated that by 2015 up to 5,400 MWe of capacity will be installed in the USA. This assumes conventional hydrothermal resources. The report assumes that known hydrothermal resources in Western USA have a potential to produce 9,000 MWe.

The most comprehensive estimates for the total US recoverable resource were produced by US Geological Survey (USGS) in the late-1970s. In 1978, USGS Circular 790 suggested that the total recoverable resource from identified geothermal prospects is roughly 23,000 MW and the total combined identified/unidentified resource base is as high as 150,000 MW. In 2008, the USGS revised its numbers to reflect lower temperature resources, apply confidence ratios to the numbers and make assertions based on 30 years of results in the field. As such, a new estimate suggested 95% confidence that identified systems can provide 3,675 MWe and 5% confidence that identified systems can provide 16,457 MWe. The new estimate suggested 95% confidence that undiscovered systems can provide 7,917 MWe and 5% confidence that undiscovered systems can provide 73,286 MWe [119].

In January of 2006, a comprehensive assessment was released by the Western Governors’ Association (WGA) in its Geothermal Task Force Report. The assessment was performed as part of the WGA’s Clean and Diversified Energy Advisory Committee (CDEAC). The report covered 11 western states

(Alaska, Arizona, California, Colorado, Hawaii, Idaho, Nevada, New Mexico, Oregon, Utah, and Washington State) and estimated that there is up to 12,558 MW of recoverable geothermal power by 2025 from identified locations available at a future market at a cost of up to 20 cents per kilowatt-hour (¢/kWh). In the near-term, WGA estimated 5,588 MW of economically developable capacity (5.3–7.9¢/kWh (with the federal production tax credit (PTC) included) by 2015 in these 11 western states [120].

Europe

Simultaneously, in Europe, Bertani [123] in Bali, 2010 reported that Europe accounts for 1,635 MWe of installed geothermal capacity with growth forecasts to 2,125 MWe by 2015. Installation of new binary power stations will increase electricity production over a wide geographical distribution in locations fueled by medium-temperature resources including nonvolcanic sources in interior Eastern and Western Europe. Further, there will be greater development in Geothermal Heat Pump (GHP) installations that can be replicated around the world. As for direct use and GHP, John Lund reported at the Geothermal Resources Council Annual Meeting in October 2010 that European nations represent 10 out of the top 15 nations in utilization of these types of installations, with Sweden in third place behind China and the USA (focused mostly on GHP). Turkey is in fourth place focused mostly on district heating (Lund and Bertani [121]).

Geothermal Resources

Evaluation of the geothermal energy reserves was compiled for the MIT publication “The Future of Geothermal Energy” [122] according to the various types of geothermal systems:

- (a) Hydrothermal convective systems
- (b) Enhanced geothermal systems (“EGS”)
- (c) Conductive sedimentary systems
- (d) Hot water produced from oil and gas fields
- (e) Geopressured systems
- (f) Magma bodies

Hydrothermal Convective Systems *Hydrothermal convective systems* to date have seen several decades

of commercial exploitation for electric power generation in about 24 countries, but their distribution worldwide is limited. The installed power capacity for such systems totaled 10,715 MW worldwide by the end of the first decade of the twenty-first century, of which 3,000 MW were in the USA. The reserve base for these systems in the USA is estimated to be in the 10,000–30,000 MW range. Technologies involved for power generation from these sources are considered mature. Data sources are in WGC [123] and GEA [124].

Enhanced Geothermal System (EGS) An *enhanced geothermal system (EGS)* implies a man-made reservoir created by hydrofracturing impermeable or very “tight” rock through wells. The creation of an EGS system is performed by injecting water in an artificially fractured reservoir well with production from another well. By using rock heated water it is possible to extract thermal energy. EGS systems are conductive systems with enhanced flow and storage capacity due to hydro fracturing. In theory, EGS can be developed anywhere in the world by drilling deep enough to encounter commercially attractive rock temperatures. However, EGS technology is still experimental and poses a series of technical challenges, such as:

- (a) Creating a pervasively fractured large rock volume
- (b) Securing commercially attractive well productivity
- (c) Minimizing the rate of cooling of the produced water with time
- (d) Minimizing the loss of the injected water through fractures
- (e) Minimizing any induced seismic effects

Sedimentary Systems Attempts are being made to develop geothermal *projects in sedimentary basins* with high heat flow (particularly in Australia). These systems are neither EGS or convective systems (due to the presence of impermeable shale layers preventing convection). No fracturing is generally needed for such systems as sedimentary rocks have intrinsic porosity and permeability. However, very deep wells are required to exploit such systems and ensure an adequate temperature level (well productivity may not prove adequate). No such systems have been

commercially exploited to date. Developing such systems should be feasible if reservoir temperatures and flow capacities are sufficiently high.

Coproduction with Oil and Gas Wells Another geothermal energy resource presently being considered for exploitation is the heat contained in the water produced from *deep oil and gas wells*. Here the hot water may be coproduced with petroleum production, from existing or abandoned oil or gas wells. While there are no significant challenges to exploiting this energy resource, the cost of this power may not always be attractive due to the relatively low temperature and low production rate of the water.

“Geopressured” Systems “*Geopressured*” systems are very restricted geothermal energy resources. These systems are confined sedimentary reservoirs with pressures greatly higher than the local hydrostatic pressure. The high pressure in such systems may allow the exploitation of the kinetic energy of the water produced in addition to its thermal energy. Furthermore, due to the high pressure, such a system may contain attractive amounts of methane gas dissolved in the water which may be used to generate electric power in a gas engine. Therefore, an ideal geopressured well can provide thermal, kinetic and gas-derived energy. No commercial geopressured project has been developed to date and there are several technical challenges to making this energy source commercial.

Magma Bodies Exploitation of geothermal energy directly from a magma body is theoretically possible but faces many technical challenges.

The US reserves of the various geothermal systems discussed above are summarized in Table 1–1 of the MIT report [122].

Of the six basic types of geothermal energy in the USA, the potential from an EGS resource system is three orders of magnitude higher than the other types combined. For additional views see [125, 126]. This conclusion also applies for the rest of the globe.

US View to 2050

Although, as described above, hydrothermal resources retain significant potential, the potential is limited

mostly to the Western USA with a smaller contribution possible from coproduced and geopressured systems from oil- and gas-producing states, such as the Gulf Coast. EGS represents a more widely distributed resource base, requiring substantial investment in R&D. An 18-member assessment panel assembled in September 2005 evaluated the technical and economic feasibility of EGS becoming a major supplier of primary energy for US baseload generation capacity by 2050. The MIT report was rediscussed in the DOE Workshop on June 7, 2007 [127], with an intention to recommend DOE action items.

The questions raised were:

1. What is the quality, grade, and distribution of the EGS resource nationally?
2. What is still to be done technically to achieve complete EGS system feasibility?
3. What are the key technical and economic issues that must be resolved for EGS to have national impact in US energy supply by 2050?

The primary goal was to provide an in-depth evaluation of EGS as a major primary energy supplier to the USA. The secondary goal was to provide a framework for informing policy makers of R&D support and policies needed for EGS to have a major impact.

Major impact was defined as enabling 100,000 MWe of an economically viable EGS resource online or as a true reserve by 2050.

Findings were:

1. Large, indigenous, accessible base load power resource – extractable amount of energy that could be recovered is not limited by resource size. EGS can sustain production of $\geq 100,000$ MWe of base load electric power.
2. Fits portfolio of sustainable RE options – EGS complements the DOE’s RE portfolio and does not hamper the growth of solar, biomass, and wind in their most appropriate domains.
3. Scalable and environmentally friendly –EGS stations have small foot prints and low carbon-free emissions. The stations are inherently modular making them easily scalable from 1+ to 50+ MWe size individual stations, grouping to large base load facilities $>1,000$ MWe.

4. Technically feasible – much progress in 30+ years of testing worldwide, the major elements of the technology to capture and extract EGS are already in place. Remaining key issue is to establish inter-well connectivity at commercial production rates – only a factor of 2–3 greater than current levels.
5. Economic projections – favorable for high grade areas now with a credible learning path to provide competitive energy from mid-and low-grade resources.
6. Deployment costs low – a modest investment of US\$300–US\$400 million over 15 years would demonstrate commercial scale EGS technology at several US field sites to reduce risks for private investment and enable the development of 100,000 MWe.
7. Supporting research costs are reasonable – in comparison to other large impact alternative energy programs supported by the US government.

The financial support recommends investing a total of US\$600–US\$800 million for deployment assistance, research and development over 15 years. This is an average of US\$50 M/year.

Refer to the EERE website <http://www1.eere.energy.gov/geothermal/> [128] for a follow-up of the DOE Geothermal Technologies Program.

Europe View to 2050

At the Offenburg conference, 2009, Bertani [129], who based his observations largely on Fridleifsson et al. [130], saw a linear increase in direct use of geothermal resources for space heating, greenhouses, etc. Simultaneously the increase of GHP including power generation grew exponentially from about 200,000 TJ/year expected in 2010 to 900,000 TJ/year expected in 2020 to above 4×10^6 TJ/year in 2050. For comparison, the world energy consumption is presently about 420 EJ/year. (1 EJ = 1,018 J, 1 TJ = 1,012 J).

Following the US evaluation, a committee of the Council of Europe met to handle the issue – Geothermal Energy: a solution for the future? A motion for a resolution is found in Doc: 11740 [131].

Presented European and world information proposed that the assembly should focus on geothermal energy and its potential contribution to clean and sustainable energy systems in Europe.

A more in-depth survey followed and was discussed by the Council of Europe in May 2010 as reported in Doc. 12249 [132]. A report of European and world geothermal data was presented at the meeting and summarized in Doc. 12249. The report handles technical data while also noting connected hurdles, technical barriers and seismic problems. Examples of such occurrences were given including Landau, Staufen-im-Brisgau in Germany and a Soultz-type geothermal project launched on a commercial basis in Basel, Switzerland in 2006. This project ceased drilling after the inhabitants reported mini earthquakes around the project site.

The draft resolution detailed the data, stressed the advantages of geothermal energy in urban heating, electric generation and positive impact on the environment. Legislative issues and financial risks were also covered. Resolution No. 9 or the “to-do” recommendations are listed below (from the original):

- 9.1 Foster the development of geothermal energy operations in their national energy strategies.
- 9.2 Encourage the use of geothermal energy in all its forms, particularly locally.
- 9.3 Encourage international cooperation in the transfer of technology and the financing of geothermal development.
- 9.4 Increase realization and awareness among the general public and potential investors of the advantages of geothermal technologies for a sustainable energy infrastructure.
- 9.5 Take the necessary steps to set up strategic research programs and encourage the exploitation of geothermal energy resources.
- 9.6 Foster the introduction of financing and insurance schemes for exploration.
- 9.7 Encourage the setting up of transfrontier cooperation schemes to finance surface measurements and test drillings.
- 9.8 Introduce a European training and professional development framework.
- 9.9 Draw up a map of geothermal energy resources at the European level within the framework of cooperation between the geological research bodies of each country.

For additional information on European data on future plan and research go to [133]. The future of

geothermal development European Geothermal Energy Council (EGEC) Projections www.egec.org, 2010 and [134] European Commission Research – Future prospects, hurdles: http://ec.europa.eu/research/energy/eu/research/geothermal/background/index_en.htm

Global View

Renewable energy including geothermal energy plays an important role in future global policies. In his book, Plan B 4.0 “Mobilization to save the Earth,” Brown [135] accounted for the existing and potential geothermal resources in the section on “[Choosing the Energy Conversion Systems](#),” which deals with renewable energy. Besides the energy use for power production he details direct usage for domestic heating, heat pumps for heating and cooling. In Germany alone he reports there are 130,000 operating heat pumps with 25,000 being added annually. Leaders for direct use of the above include:

- District heating: Iceland, Hungary, France, and China.
- Pools and spas in Iceland, France, and Japan.
- Greenhouses in Russia, Hungary, Iceland, and USA.
- Aquaculture in China, Israel, and the USA.

The previously mentioned GEA report 2010 [124] summarizes the existing portfolio of power stations and gives a prospect for increase until 2015. Highlights are:

According to the International Geothermal Association in 2005, there was 8,933 MW of installed power capacity in 24 countries, generating 55,709 GWh of green power per year. IGA reports in 2010 that there is 10,715 MW online generating 67,246 GWh. This represents a 20% increase in online geothermal power between 2005 and 2010. IGA projects this will grow to 18,500 MW by 2015, as based on the large number of projects under consideration (Bertani [123]).

Countries which experienced significant increases in installed capacity between 2005 and 2010 included the USA, Indonesia, Iceland, New Zealand and Turkey. These nations expect to have a significant increase by 2015. However, other nations expect to increase generation significantly during that time including Kenya, Russia, the Central American nations of El Salvador, Guatemala, Nicaragua and the South American nation of Chile, which currently has 0 MW of installed capacity (Bertani [123]).

While on-line power increased 20% between 2005 and 2010, countries with projects under development grew at a much faster pace. GEA reported in 2007 that 46 countries were considering geothermal power development. In 2010, an updated report acknowledged 70 countries with projects under development or active consideration, a 52% increase since 2007 (GEA [124]).

Projects under development grew most dramatically in two regions of the world, Europe and Africa. Ten countries in Europe were listed as having geothermal projects under development in 2007, and this has more than doubled in 2010 to 24 countries. Six countries in Africa were identified in 2007 and in 2010, 11 were found to be actively considering geothermal power. It would appear that bodies such as ARGeo and the European Bank for Reconstruction and Development's geothermal initiatives are having a considerable beneficial effect.

However, despite these growth trends, potential of geothermal resources to provide clean energy appears to be underestimated. In 1999, GEA prepared a report that examined geothermal power potential internationally. The report showed that in the vast majority of countries the estimated potential remains undeveloped and largely untapped, even assuming the lowest projections for geothermal resource potential. Moreover, the number of countries with geothermal power potential still not developing their resources is still high. Of the 39 countries identified in 1999 as having the potential to meet 100% of their electricity needs through domestic geothermal resources, significant power production had been developed in only nine – Costa Rica, El Salvador, Guatemala, Iceland, Indonesia, Kenya, Nicaragua, Papua New Guinea and the Philippines. However, this report identified projects under consideration in another 14 of these countries. (For a list of countries identified in the 1999 GEA report which could be 100% geothermal powered, see the Appendix in [124].)

The underlying trend of geothermal power expansion is complemented by the development of projects in entirely new areas. It is interesting to note that there are 24 countries identified with geothermal power projects under development not included in the GEA 1999 study. Most of these countries are in Europe and are accessing resources with new technology developments that allow development of lower-temperature resources. In addition, EGS technologies, or enhanced geothermal systems,

are being developed in a number of countries including Australia, France, Germany, the UK and the USA.

The trends in both the number of new countries developing geothermal energy and the total of new megawatts of power capacity under development appear to continue a growth trend showing a clear reverse from the slowdowns in international markets as seen in the late 1990s. Supported by the development of low-temperature power on the one side and EGS technologies on the other side, the geothermal market appears to be expanding to encompass most of the world's potential geothermal sites.

The report indicates that national and international policies, as well as financial support, are key in realizing the potential for successful geothermal development.

Additional GEA Observations:

- In 2010, global geothermal development is partly being driven by a number of regional institutions which, in addition to financing geothermal projects, are enhancing regional cooperation within an emerging renewable energy sector. Examples include the African Rift Geothermal Energy Development Facility (ARGeo), which underwrites drilling risks in six African nations and is backed by UNEP, the World Bank, and the geothermal initiatives of the European Bank for Reconstruction and Development supported by European Union climate policies.
- Geothermal development appears to be increasingly supported by a global financial market. A growing number of countries, including Australia, China, Germany, Iceland, Italy, Japan, and the USA, are facilitating geothermal development projects around the world. Forms of support other than financing, including technology sharing, training, and geological surveys are also being endorsed by outside governments.
- The growth in geothermal projects under consideration or in development is in part attributable to international and multilateral support for development in new areas. The ongoing question is whether that support will be sustained over time and be adequate to address risks involved in geothermal project development. For example, geothermal resources are abundant in East Africa and support for resource assessment has helped spur interest in project development in several countries. But, new projects will have high associated costs and risk factors. Sustained support for development at this crucial stage is essential to achieving expanded use of geothermal energy in this and other developing areas.
- Geothermal development appears to be trending beyond traditional hydrothermal reserves prevalent along the Pacific Ring of Fire. Lower temperature power systems and EGS technology are allowing a growing and diversified collection of countries to actively pursue geothermal development in areas previously assumed to have little exploitable resource. This is especially true among European countries, notably France, Germany, Latvia, and the UK, all of which are currently exploring and developing local resources by employing EGS. These developments are supported by government policies (such as feed-in tariffs), which make higher-risk and higher-cost projects more feasible. These policies are typically components of broader climate initiatives.
- District heating and direct use of geothermal applications appear to be progressively more commonplace in many countries and are being emphasized in a number of national renewable energy policies as effective measures for curbing greenhouse gas emissions.
- Around the world, villages and tribes are looking to geothermal as a way to utilize land and become energy independent. Warm Springs Indian Reservation in Oregon, the Northwestern Band of Shoshone Nation in Idaho and Utah, and the Jemez Pueblo in New Mexico have all shown interest in developing geothermal energy. Additionally, the Pyramid Lake Paiute Tribe in Nevada is actively developing its geothermal resources and was recently awarded funding from the US DOE. In New Zealand, the Te Arawa Iwi is examining the possibility of geothermal power on Maori land in Rotorua. In The Philippines, 9 of 11 ancestral domain areas consented to the Kalinga geothermal exploration project. A geothermal station is expected to open in the small settlement of Innamincka, Australia, in early 2012.

A country-by-country assessment (both present and forecasted for 2015) is summarized in the WGC report by Bertani [123] and in the GEA report [124] which is arranged by continents considering natural geothermal data and national policies.

Acknowledgments

The author would like to acknowledge with appreciation the valuable contributions of Dr. Uriyel Fisher and Mr. Mike Kanowitz in preparation and typing of the manuscript as well as for the special attention to accuracy and detail.

Bibliography

1. U.S. Department of Energy (2008) The price of geothermal power. Geothermal Tomorrow 2008, pp 19–21. http://www1.eere.energy.gov/geothermal/pdfs/geothermal_tomorrow_2008.pdf
2. California Energy Commission (2007) Comparative costs of California Central Station Electricity Generation Technologies. Document #: CEC-200-2007-011-SF. <http://www.energy.ca.gov/2007publications4CEC-200-2007-011/CEC-200-2007-011-SF.pdf>
3. Emerging Energy Research (2009) Global geothermal markets and strategies: 2009–2020. Section 3: Geothermal technology and cost trends. <http://www.emerging-energy.com/Content/Document-Details/4/Global-Geothermal-Markets-and-Strategies-20092020/280.aspx>
4. Glacier Partners (2009) Geothermal economics 101 – economics of a 35 MW binary cycle geothermal plant. <http://www.glacierny.com/geothermal.php>
5. Public Utilities Commission of Nevada (PUCN) (2010) Docket # 10-02009 “Application of Nevada Power Company d/b/a NV energy for approval of its 2010–2029 triennial integrated resource plan.” NV Energy PPA Pricing Info. <http://pucweb1.state.nv.us/pucn/DktInfo.aspx?Util=Electric>
6. International Energy Agency (IEA) (2010) Renewable energy essentials: geothermal. http://www.iea.org/papers/2010/Geothermal_Essentials.pdf
7. California Energy Commission (2009) Renewable energy cost of generation update. Document #: CEC-500-2009-084. <http://www.energy.ca.gov/2009publications/CEC-500-2009-084/CEC-500-2009-084.PDF>
8. Remo AR (2010) EDC to put up geothermal plants worth \$1B. Philippine Daily Inquirer. <http://business.inquirer.net/money/topstories/view/20100729-283909/EDC-to-put-up-geothermal-plants-worth-1B>
9. Kema, Inc., California Energy Commission (2009) Renewable energy cost of generation update. Document #: CEC-500-2009-084, pp 52–72; Appendix A, pp 206–215. <http://www.energy.ca.gov/2009publications/CEC-500-2009-084/CEC-500-2009-084.PDF>
10. California Public Utilities Commission (CPUC) (2010) RPS project status table – July update. http://www.cpuc.ca.gov/NR/rdonlyres/A5406F32-B0D0-409E-AA92-0EA79E97BECC/0/RPS_Project_Status_Table_2010_July.xls
11. Geothermal Energy Association, http://geo-energy.org/geo_basics_plant_cost.aspx
12. Cappetti G et al (2000) Italy country update report 1995–1999. In: Proceedings world geothermal congress 2000, Kyushu – Tohoku, 28 May–10 June 2000
13. Calpine Geothermal, <http://www.ncpageo.com/about.htm>
14. The Geysers, <http://www.geysers.com/>
15. Rhinehart JS (1980) Geysers and geothermal energy. Springer, New York
16. Kestin J (1980) Sourcebook on the production of electricity from geothermal energy. DOE, RA/4051-1. US Government Printing Office, Washington, DC
17. Dipippo R (1998) Geothermal power systems, Section 8.2. In: Elliot TC, Chen K, Swanecamp RC (eds) Standard handbook of power plant engineering, 2nd edn. McGraw-Hill, New York
18. Chacko J et al (1998) Gulf coast geopressured-geothermal program summary report compilation. U.S. Department of Energy, Contract no. DE-FG07-95ID 13366
19. Duchane D et al (2002) Hot Dry Rock (HDR) geothermal energy research and development at Fenton Hill, New Mexico. GHC Bull 13–19
20. Lund JW (2007) Characteristics, development and utilization of geothermal resources. GHC Bull 28:1–9. <http://geoheat.oit.edu/bulletin/bull28-2/art1.pdf>
21. Dipippo R (2008) Geothermal power plants: principles, applications and case studies, Chapters 6 & 9. Elsevier
22. Ellis AJ, Mahon WAJ (1977) Chemistry and geothermal systems. Academic, New York
23. Grassiani M (1994) Advances in materials selection for geothermal power production application. In: Coutsouradis D et al (eds) Materials for advanced power engineering. Kluwer, Dordrecht, pp 1677–1684
24. Mitsubishi Jukogyo Kabushiki Kaisha (1984) Geothermal power generation, rev edn. Mitsubishi Heavy Industries, Tokyo
25. Fuji brochure (1988) Information data on geothermal power plant. Fuji Electric, Japan
26. Kestin J (1980) Available work in geothermal energy, Chapter 3. In: Sourcebook on the production of electricity from geothermal energy, RA/4051-1. US Government Printing Office, Washington, DC
27. White DE (1973) Characteristics of geothermal resources, Chapter 4. In: Kruger P, Otte C (eds) Geothermal energy: resources, production, stimulation. Stanford University Press, Stanford, pp 69–94
28. James R (1968) Pipeline transmission of steam-water mixtures for geothermal power. NZ Eng 23:55–61
29. Baumann K (1921) Some recent developments in large steam turbine practice. J Inst Electric Eng 59:565
30. http://www.thermoflow.com/ConvSteamCycle_TFX.htm
31. Keenan JH, Keyes EG, Hill PG, Moore JG (1969) Steam tables: thermodynamic properties of water including vapor, liquid, and solid phases (international edition – metric units). Wiley, New York
32. Wagner W, Kretschmar H-J (2008) International steam tables: properties of water and steam based on industrial formulation IAPWS-IF97. Springer, Berlin

33. DiPippo R (2008) Geothermal power plants: principles, applications, case studies and environmental impact, 2nd edn. Elsevier, Oxford
34. Anonymous (1989) East Mesa 18.5 MW \times 2 double flash cycle geothermal power plant. Mitsubishi Heavy Industries, Tokyo
35. DiPippo R (1998) Geothermal power systems, Sect. 8.2. In: Elliott TC, Chen K, Swanekamp RC (eds) Standard handbook of powerplant engineering, 2nd edn. McGraw-Hill, New York, pp 8.27–8.60
36. Bronicki LY (2008) Advanced power cycles for enhancing geothermal sustainability 1,000 MW deployed worldwide. In: IEEE PES general meeting, Pittsburg. McGraw-Hill, New York, pp 8.27–8.60
37. Incropera PP, DeWitt DR (1996) Fundamentals of heat and mass transfer, 4th edn. Wiley, New York
38. Krieger Z et al (1986) Cascaded power plant using low and medium temperature source fluid. US Patent 4,578,953, 01 April 1986
39. Khalifa HE, Rhodes BW (1985) Analysis of power cycles for geothermal wellhead conversion systems, EPRI AP-4070. Electric Power Research Institute, Palo Alto
40. Bliem CJ, Walrath LR (1983) Raft river binary-cycle geothermal pilot power plant final report, EGG-2208. Idaho National Engineering Laboratory, Idaho Falls
41. Tester JW, Milora SL (1976) Geothermal energy as a source of electric power. In: 16th annual symposium, New Mexico Section, American Society of Mechanical Engineering, Albuquerque, 26–27 Feb 1976
42. Reynolds WC (1979) Thermodynamic properties in SI: graphs, tables and computational equations for 40 substances. Department of Mechanical Engineering, Stanford University, Stanford
43. Gallagher JS, Linsky D, Morrison G, Levelt Sengers JMH (1987) Thermodynamic properties of a geothermal working fluid; 90% isobutane 10% isopentane, NBS Technical Note 1234. National Bureau of Standards, U.S. Government Printing Office, Washington, DC
44. Berning J et al (1988) Heber binary cycle geothermal demonstration power plant; half load testing. Special report EPRI AP-5787-SR. Electric Power Research Institute, Palo Alto
45. Demuth OJ, Bliem CJ, Mines GL, Swank WD (1975) Supercritical binary geothermal cycle experiments with mixed-hydrocarbon working fluids and a vertical, in-tube, counter-flow condenser, EGG-EP-7076. Idaho National Engineering Laboratory, Idaho Falls
46. IAPWS (2003) Giddeline on the Tabular Taylor Series Expansion (TTSE) method for calculation of thermodynamic properties of water and steam. Applied to IAPWS-95 as an example. International Association for the Properties of Water and Steam, Vejle
47. Milora SL, Tester JW (1976) Geothermal energy as a source of electric power: thermodynamic and economic criteria. MIT Press, Cambridge, MA
48. Anonymous (1997) ASHRAE handbook fundamentals, Chapter 18. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Atlanta
49. DiPippo R (1990) Geothermal power cycle selection guidelines, part 2 of geothermal information series, DCN 90-213-142-02-02. Electric Power Research Institute, Palo Alto
50. Kestin J (Ed. In Chief), DiPippo R, Khalifa HE, Ryley DJ (eds) (1980) Sourcebook on the production of electricity from geothermal energy. U.S. Department of Energy, DOE/RA/4051-1. U.S. Government Printing Office, Washington, DC
51. Tabor H, Bronicki LY (1961) Small turbines for solar energy package. In: United Nations conference on new sources of energy, Rome
52. Tabor H, Bronicki LY (1964) Establishing criteria for fluids for small vapor turbines, SAE Power 931C
53. Tabor H, Bronicki LY (1962) Vapor turbines. US Patent 3,040,528, 26 June 1962
54. Lazalde-Crabtree H (1984) Design approach of steam-water separators and steam dryers for geothermal applications. Geotherm Resour Counc Bull 13(8):11–20
55. http://www.htri.net/articles/htri_xchanger_suite
56. McKetta JJ (1992) Piping design handbook. Marcel Dekker, New York (Originally Encyclopedia of chemical processing and design, 1991)
57. Salomon L (1999) Two-phase flow in complex systems. Wiley, New York
58. Cheremisinoff NP (ed) (1986) Encyclopedia of fluid mechanics, vol 3, Gas-liquid flows. Gulf, Houston
59. Wallis GB (1969) One-dimensional two-phase flow. McGraw-Hill, New York
60. Holm A, Blodgett L, Jennejohn D, Gawell K (2010) Geothermal energy: international market update. GEA, Washington, DC
61. Bronicki LY (1995) Innovative geothermal power plants, fifteen years of experience. In: World geothermal conference, Florence
62. Blaydes PE (1994) Environmental advantages of the binary power plants can enhance development opportunities. Trans Geotherm Resour Counc 18:121–125
63. Flynn T (1997) Geothermal sustainability, heat utilization and advanced Organic Rankine cycle. Trans Geotherm Resour Counc 26(9):224–229. Sept/Oct 1997
64. Sanyal SK (2005) Sustainability and renewability of geothermal power capacity. In: Proceedings world geothermal congress 2005, Antalya, 24–29 April 2005
65. Wilson SS, Radwan MS (1977) Appropriate thermodynamics for heat engine analysis & design. Int J Mech Eng 5:68–82
66. History of geothermal energy, <http://www1.eere.energy.gov/geothermal/history.html>
67. LeConte JL (1855) Account of some volcanic springs in the desert of the Colorado, in Southern California. Am J Sci Art Second Ser 19:1–6
68. Lombard GL (1978) Operational experience at the San Diego gas & electric ERDA Niland geothermal loop experimental

- facility. In: Proceedings of the EPRI annual geothermal program project review and workshops EPRI ER-660-SR. Electric Power Research Institute, Palo Alto, pp 3-11-3-16
69. Featherstone J, Butler S, Bonham E (1995) Comparison of crystallizer reactor clarifier and pH mod technologies at the Salton Sea geothermal field. In: Proceedings world geothermal congress 1995, vol 4. International Geothermal Association, pp 2391-2396
70. Anonymous (2001) CalEnergy Company, Inc., U.S.A. Salton Sea Unit 5 Geothermal Power-Plant 1 X 58.32 MW, Brochure GEC 82-14. Fuji Electric, Tokyo
71. Gallup DL (1993) Control of salt precipitation from geothermal brine. US Patent 5,256,301, 26 Oct 1993
72. Kits KR (1997) pH modification of geothermal brine with sulfur-containing acid. US Patent 5,656,172, 12 Aug 1997
73. DiPippo R (2008) Geothermal power plants: principles, applications, case studies and environmental impact, 2nd edn. Elsevier, Oxford, p 225
74. Clutter TJ (2000) Mining economic benefits from geothermal brine. *GHC Bull* 21:1-3
75. Cal Energy Publication, <http://www.calenergy.com/projects2d.aspx>
76. Krieger Z, Moritz A (1986) Cascaded power plant using low and medium temperature source fluid. US Patent 4,578,953, 01 April 1986
77. Tabor H, Bronicki L (1962) Vapor turbines. US Patent 3,040,528, 26 June 1962
78. Legman H, Sullivan P (2003) The 30 MW Rotokawa I geothermal project five years of operation, IGC -20034164-2003
79. Lienau PJ (1996) Sudurnes Regional Heating Corporation, Geo-Heat Center. *GHC Bull* 17(4):14-16
80. Bronicki LY (1981) Practical experience and potential of organic vapor turbines for onsite electricity production from small local energy resources. UNITAR, Los Angeles
81. Kaplan U (1997) Method and apparatus for producing power using geothermal fluid. US Patent 5,664,419, 9 Sept 1997
82. Bronicki LY (1985) Geothermal power plant and method utilizing the same. US Patent 4,542,625, 24 Sept 1985
83. Hinrichs TC, Dambly BW (1980) East mesa magmamax power process geothermal generating plant: a preliminary analysis, EPRI TC-80-907. In: Proceedings fourth annual geothermal conference and workshop, Electric Power Research Institute, Palo Alto, pp 5-1-5-14
84. Hinrichs TC (1984) Magmamax power plant - success at east mesa, EPRI AP-3686. In: Proceedings eighth annual geothermal conference and workshop, Electric Power Research Institute, Palo Alto, pp 6-21-6-30
85. Ferdinand C (1850) <http://www.probrewer.com/resources/refrigeration/history.php>
86. Kalina A (1986) Method and apparatus for implementing a thermodynamic cycle using a fluid of changing concentration. US Patent 4,586,340, 6 May 1986
87. Bliem CJ, Mines GL (1991) Advanced binary performance power plants: limits of performance, EGG-EP-9207. Idaho National Engineering Laboratory, Idaho Falls
88. DiPippo R (2002) Second low basis for efficient power generation from industrial wasteheat. Report made for Ormat International Inc, Sparks NV
89. Bombarda P et al (2010) Heat recovery from diesel engines: a thermodynamic comparison between Kalina and ORC cycles. *Appl Therm Eng* 30:212-219
90. Gulf Coast Geopressured-Geothermal program summary report compilation work performed under U.S. Department of Energy Contract No. DE-FG07-95ID 13366
91. Griggs J (2004) A re-evaluation of geopressured-geothermal aquifers as an energy resource. Master's thesis, Louisiana State University, Craft and Hawkins Department of Petroleum Engineering
92. Nitschke GS, Harris JA (1991) Production of fresh water and power from geopressured-geothermal reservoirs. In: Indirect solar, geothermal and nuclear energy. Nova Science, New York
93. Árpási M, Lorberer Á, Pap S (2000) High pressure and temperature (geopressured) geothermal reservoirs in Hungary. In: Proceedings world geothermal congress 2000, International Geothermal Association, pp 2511-2514
94. He L, Xiong L (2000) Extensional model for the formation of geopressured geothermal resources in the Yinggehai Basin, South China Sea. In: Proceedings world geothermal congress 2000, International Geothermal Association, pp 1211-1216
95. Anonymous (1924) Recent developments in the utilization of the earth's heat. *Mech Eng* 46(8):448-449
96. DiPippo R (1978) An analysis of an early hybrid fossil-geothermal power plant proposal. *Geotherm Energy Magazin* 6(3):31-36
97. Janes J (1984) Evaluation of a superheater enhanced geothermal steam power plant in the geysers area, Rep. P700-84-003. California Energy Commission, Sitting and Environmental Division
98. Kestin J, DiPippo R, Khalifa HE (1978) Hybrid geothermal-fossil power plants. *Mech Eng* 100:28-35
99. DiPippo R, Khalifa HE, Correia RJ, Kestin J (1979) Fossil superheating in geothermal steam power plants. *Geotherm Energy Magazin* 7(1):17-23
100. Chang I, Williams JR (1985) Thermodynamic analysis of a geopressured geothermal hybrid wellhead power system. Final report. DOE/NV/10355-1
101. Future of Geothermal Energy, MIT report. http://geothermal.inel.gov/publications/future_of_geothermal_energy.pdf
102. Vuataz FD (2004) Hijiori hot dry rock project, northern Japan. Swiss Deep Heat Mining Project, Steinmaur. <http://www.dhm.ch/imaH00hijiori.html>
103. New Energy and Industrial Technology Development Organization (2004) Development of a hot dry rock power generation system, New Energy and Industrial Technology Development Organization, Kanagawa. <http://www.nedo.go.jp/chinetsu/hdr/indexe.htm>
104. Vuataz FD (2004) Ogachi hot dry rock project, northern Japan, June 2000. Swiss Deep Heat Mining Project. Steinmaur. <http://www.dhm.ch/imaOG00ogachi.html>

105. Kruger P, Karasawa H, Tenma N, Kitano K (2000) Analysis of heat extraction from the Hijiori and Ogachi HDR geothermal resources in Japan. In: Proceedings world geothermal congress 2000, International Geothermal Association, pp 2677–2682
106. Brown DW (1996) 1995 reservoir flow testing at Fenton Hill, New Mexico. In: Proceeding of the 3rd international HDR forum, Santa Fe, pp 34–37
107. MacDonald P, Stedman A, Symons G (1992) The UK geothermal HDR R&D programme. In: Proceedings seventeenth workshop on geothermal reservoir engineering, SGP-TR-141, Stanford University, Stanford, 29–31 Jan 1992
108. Opportunities in the upper Rhine Valley, http://www.energy-base.org/fileadmin/media/regioner/docs/geothermal-energy-rhine-valley_FINAL.pdf
109. Baria R, Baumgaertner J, Teza D, Michelet S (2005) Reservoir stimulation and testing techniques for EGS systems (Soultz). Presented at EGS workshop, Massachusetts Institute of Technology, Cambridge, MA, 10 Nov 2005
110. Smith IK (1993) Development of the trilateral flash cycle system. Part 1: fundamental considerations. *Proc Inst Mech Eng A* 207(A3):179–194
111. Austin AL, Lundberg AW (1978) The LLL geothermal energy program: a status report on the development of the total-flow concept, UCRL-500-77. Lawrence Livermore Laboratory, Livermore
112. McKay R (1982) Helical screw expander evaluation project: final report, DOE/ET-28329-1, JPL Pub. 82-5. Jet Propulsion Laboratory, Pasadena
113. Carey B (1983) Total flow power generation from geothermal resources using a helical screw expander. In: Proceeding of the 5th New Zealand geothermal workshop, pp 127–132
114. Cerini DJ, Record J (1983) Rotary separator turbine performance and endurance test results. In: Proceedings of the seventh annual geothermal conference and workshop, EPRI AP-3271. Electric Power Research Institute, Palo Alto, pp 5-75–5-86
115. Sprankle R (1986) Helical screw expander power plant model 76-1 test result analysis. In: Proceedings of a Topical meeting on Small scale geothermal power plants and geothermal power projects, 12–13 Feb 1986. Hydrothermal Power Co., Reno Nevada, pp 39–58
116. Hughes EE (1986) Summary report: rotary separator turbine. Final report, EPRI AP-4718. Electric Power Research Institute, Palo Alto
117. King Hubbert M (1975) Survey of World Energy Resources. In: Symposium on energy sources for the future, Oak Ridge, 7–25 July 1975
118. Tester JW, Milora SL (1975) Geothermal energy. In: Symposium on energy sources for the future, Oak Ridge, 7–25 July 1975
119. Geothermal Energy Association, Geothermal basics potential use. <http://www.geo-energy.org/PotentialUse.aspx>
120. Western Governors Association (WGA) (2006) Geothermal task force report. <http://www.westgov.org/wga/initiatives/cdeac/Geothermal-full.pdf>
121. Lund JW, Bertani R (2010) Worldwide geothermal utilization 2010. In: Geothermal resources council annual meeting 2010, Sacramento, 25–27 Oct 2010. GRC transactions, vol 34, pp 195–198
122. The future of geothermal energy, impact of Enhanced Geothermal Systems (EGS) on the United States in the 21st century (2006) MIT Press. http://www1.eere.energy.gov/geothermal/egs_technology.html
123. Bertani R (2010) Geothermal power generation in the world 2005–2010 update report. In: Proceedings world geothermal congress 2010, Bali, 25–29 April 2010
124. Holm A, Blodgett L, Jennejohn D, Gawell K (2010) Geothermal energy: international market update. Geothermal Energy Association, Washington, DC
125. Sanyal SK (2010) Future of geothermal energy. In: Proceedings, thirty-fifth workshop on geothermal reservoir engineering, SGP-TR-188, Stanford University, Stanford, 1–3 Feb 2010
126. Johanson TB, Goldenberg J (2004) World energy assessment overview 2004 update. UNDP 2005, <http://www.undp.org/energy/weaover2004.htm>
127. Tester J, DiPippo R (2007) The future of geothermal energy structure and outcome of the analysis. In: Presentation at the DOE geothermal program workshop, Washington, DC, 7 June 2007
128. DOE Geothermal Technologies program use EERE website. <http://www1.eere.energy.gov/geothermal/>
129. Bertani R (2009) Long term projection of geothermal electricity development in the world. In: Geotherm 2009 expo & congress, Offenburg, 5–6 March 2009. <http://www.iea-gia.org/documents/LongTermGeothermElecDevelopWorldBertanioffenburg23Feb09.pdf>
130. Fridleifsson IB, Bertani R, Huenges E, Lund JW, Ragnarsson A, Ryback L (2008) The possible role and contribution of geothermal energy to the mitigation of climate change. In: IPCC scoping meeting on renewable energy sources, Luebeck, 21–25 Jan 2008
131. Council of Europe, Doc. 11740, 10 October 2008 Geothermal Energy: a solution for the future? Motion for a resolution presented by Mrs Bjarnadottir and others
132. Council of Europe, Doc. 12249, 6 May 2010. Geothermal energy – a local answer to a hot topic? Report by: Mr René ROUQUET, France, Committee on the Environment, Agriculture and Local and Regional Affairs. http://assembly.coe.int/ASP/Doc/DocListingDetails_E.asp?DocID=13060
133. EGEC (2010) The future of geothermal development, European Geothermal Energy Council (EGEC) projections. www.geg.org
134. European Commission Research, Future prospects, hurdles. http://ec.europa.eu/research/energy/eu/research/geothermal/background/index_en.htm
135. Brown LR (2009) Stabilizing climate: shifting to renewable energy, Chapter 5. In: Plan B 4.0: mobilizing to save civilization. W.W. Norton, New York. www.earthpolicy.org/index.php?/books/pb4
136. Hayes L (2011) Demonstration of a variable phase turbine. <http://energent.net>

Geothermal Power Economics

SUBIR K. SANYAL

GeothermEx, Inc, Richmond, CA, USA

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Factors That Determine Geothermal Power Cost

Estimating Levelized Power Cost

Sensitivity of Levelized Power Cost

Concluding Remarks

Future Directions

Bibliography

Glossary

Capital cost Capital costs are the one-time costs incurred on project acquisition, drilling, construction, and equipment needed to bring a project to a commercially operable status.

Levelized power cost The present value of the total cost of developing and operating a geothermal power plant over its economic life divided by the total power generated over the same period, costs being levelized in real dollars (i.e., adjusted to remove the impact of inflation).

Make-up well cost Cost of drilling “make-up” wells as needed during project operation.

Operations and maintenance (O&M) Cost Those expenses used for the day-to-day operation of a power facility. The major categories include personnel, general and administrative, insurance, supplies and services, well maintenance, and equipment maintenance costs.

Power capacity The maximum output of power from a power plant, commonly expressed in megawatts (MW).

Definition of the Subject and Its Importance

Geothermal power is the rate of extraction of geothermal energy, whether expressed as heat energy or equivalent electrical energy, and is expressed as Watt or an equivalent unit. The extraction of geothermal energy,

and therefore geothermal power capacity, is dependent not only on the technological barriers to this energy extraction but also on the economic barriers. Power generation from geothermal energy, therefore, requires consideration of the economics of geothermal power. This entry considers power cost as the main economic criterion rather than the power price or project profitability because, unlike price or profitability, cost is substantially independent of the corporate culture of the developer and operator, financing mechanism, local market forces, and government policies. The most comprehensive measure of the geothermal power cost is “levelized” power cost, expressed typically as cents per kW-hour power generated over the life of a power plant.

Introduction

The power cost considered here is “levelized” cost (¢/kWh) over the project life, defined here as the cumulative present value of all future costs including annual payments for amortized capital in real dollars (adjusted for inflation) divided by the cumulative power generated [1]. The initial capital cost is amortized over a period of 30 years; make-up well drilling cost is not capitalized and is considered an operating expense. The capital cost includes the cost of money (i.e., the cumulative future interest payments discounted for inflation) but does not include any transmission line cost or any unusually site-specific costs of regulatory compliance or environmental impact mitigation.

Cost calculations in this entry ignore any royalty burden, tax liability, or tax credit. The values of economic parameters assumed in this entry reflect the setting in the USA as of year 2005. However, levelized cost of geothermal power has approximately doubled between 2005 and 2010 because of large increases in drilling cost and commodity prices. Even so, the conclusions arrived at should still be applicable at least qualitatively to geothermal power projects worldwide. In the debate over the relative virtues of various forms of renewable energy, power cost is an objective criterion that should favor geothermal; yet there is considerable difference of opinion as to what it truly is and can be.

The analysis, extracted from Ref. [1], considers a power capacity range of 5–150 MW with 50 MW as

the “base case.” Power cost consist of three components: (1) capital cost component (including cost of money), (2) operations and maintenance (O&M) cost component (not counting debt service, which is included under the capital cost component), and (3) make-up well drilling cost component.

Factors That Determine Geothermal Power Cost

These factors can be grouped into four categories: (1) economy of scale, (2) well productivity characteristics, (3) development and operational options, and (4) macroeconomic climate. In general, economy of scale allows both unit capital cost (in US dollars per kilowatts installed) and unit O&M cost (in ¢/kWh) to decline with increasing installed capacity. Based on the data presented by Entingh and McVeigh [2], the unit capital cost (as of 2005) is estimated to vary from \$1,600/kW to \$2,500/kW depending on project size and other project-specific criteria. For the smallest project size of 5 MW considered here, the author has assumed a unit capital cost of \$2,500/kW and for the largest considered project size of 150 MW a cost of \$1,600/kW. A permissive assumption has been made that within the above range of values, unit capital cost declines exponentially with plant capacity. This assumption leads to the following correlation between unit capital cost in \$/kW (c_d) and plant capacity in kW (P):

$$c_d = 2500e^{-0.003(P-5)} \quad (1)$$

For the 50 MW base case, the unit capital cost is estimated from Eq. 1 at \$2,184/kW. GeothermEx's experience shows the representative unit O&M cost approximately ranged from 2.0 ¢/kWh for a 5 MW plant to 1.4 ¢/kWh for a 150 MW plant in 2005. Assuming an exponential decline in unit O&M cost in ¢/kWh (c_o) with plant capacity in kW (P):

$$c_o = 2.0e^{-0.0025(P-5)} \quad (2)$$

For the 50 MW base case, the unit O&M cost is estimated from Eq. 2 at 1.79 ¢/kWh.

Well productivity characteristics affect geothermal power cost in mainly two ways:

1. If well productivity is higher, fewer wells are needed to supply a plant, thus reducing power cost.

2. A higher rate of decline in well productivity with time calls for more make-up well drilling, and therefore, leads to higher power cost.

For the purposes of this entry, an average initial productivity of 5 MW per well was assumed; this is a typical value. Geothermal wells generally undergo “harmonic” decline in well productivity with time [3]:

$$W = \frac{W_i}{1 + D_i t} \quad (3)$$

where W_i is initial productivity, D_i is initial annual decline rate in productivity, and W is productivity in year t . The harmonic decline trend implies a decline rate that slows down with time, the annual decline rate (D) in productivity in year t being given by [3]:

$$D = \frac{D_i}{1 + D_i t} \quad (4)$$

If the total production rate from a field is small enough to be entirely compensated by natural recharge or if only a small fraction of the productive reservoir is being exploited, the decline rate in well productivity would be insensitive to increases in plant capacity. These situations are much less common. In most cases, decline rate increases with increasing installed capacity. This sensitivity of productivity decline to installed capacity is too site-specific to be quantified by a generally applicable correlation. Nevertheless, Sanyal et al. [4] attempted an approximate formulation:

$$D'_i = \left(\frac{W'_i}{W_i} \right) \left(\frac{\ln W'_i}{\ln W_i} \right) D_i, \quad (5)$$

where D_i is initial annual harmonic decline rate when total production rate is W_i and D'_i is initial annual harmonic decline rate when total production rate is changed to W'_i . Assuming a typical initial harmonic decline rate of 5% per year for the 50 MW base case, the initial annual harmonic decline rate for any other plant capacity was estimated from Eq. 5.

There are certain resource development and operational options that affect power cost. The developer of a geothermal project has the option to size the power

plant while the operator of the project has the option either to allow generation to decline with time or to maintain generation by make-up well drilling; the operator can also run the plant beyond its amortized life. The sensitivity of power cost to these intertwined options has been studied in this entry. The resource development option has been considered by varying the plant capacity within the range of 5–150 MW. The operational option has been considered by assuming make-up well drilling for various periods of time following plant start-up, and scenarios of plant operation both up to and beyond the amortization period.

While the unit capital cost for a given plant capacity, as given by Eq. 1, includes initial drilling cost, the unit O&M cost given by Eq. 2 does not include make-up well drilling cost. In order to estimate the make-up well drilling cost as a function of time, it is necessary to estimate first the initial number of wells required for a given plant capacity. This estimate was based on a typical initial productivity of 5 MW per well plus the customary need for at least one standby well and a minimum of 10% reserve production capacity at all times. With the above assumptions, it follows that the installed plant capacity can be maintained without any make-up well drilling for up to t_c years following plant start-up, as given by:

$$t_c = \frac{1}{D_i} \left[\frac{W_i N_{wi}}{(1 + r/100)^P} - 1 \right], \quad (6)$$

where D_i is initial annual harmonic decline rate, W_i is initial productivity per well (MW), N_{wi} is initial number of wells (including at least one standby well), P is plant capacity (MW), and r is minimum production capacity reserve required (%).

Estimating Levelized Power Cost

Figure 1 shows the schematic generation and make-up well drilling histories of a typical power project. Generation can be maintained without make-up well drilling up to year t_c as given by Eq. 6. Then generation is maintained by make-up well drilling up to year t_d in response to decline in well productivity according to Eq. 3, the initial annual harmonic decline rate being given by Eq. 5. After year t_d no make-up well is drilled and generation is allowed to decline as per Eqs. 3 and 5.

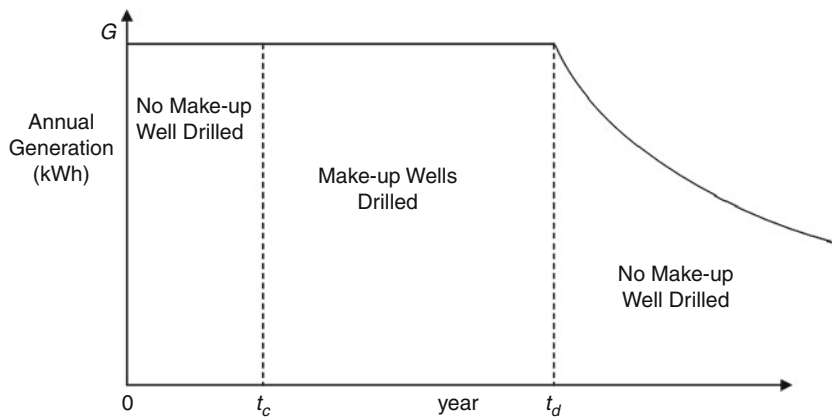
Given the generation and make-up well drilling histories represented in Fig. 1, the levelized cost of geothermal power (\bar{c}) in ¢/kWh is given by [1]:

$$\begin{aligned} \bar{c} = & \frac{100D(t_d)}{G\{D(t_d)t_d + \ln[1 + D(t_d)(n - t_d)]\}} \\ & \cdot \left\{ \frac{iC(1+i)^n}{(1+i)^n - 1} \right\} \left\{ \frac{(1+I)^n - 1}{I(1+I)^{n-1}} \right\} + c_{ov} + \left(\frac{t_d}{n} \right) c_{ofi}, \\ & + \frac{c_{ofi}}{n} \left\{ (n - t_d) + \frac{D(t_d)}{2} (n - t_d)^2 \right\} \\ & + \frac{100C_{wi}N_{wi}D(t_d)D(t_c)(t_d - t_c)}{G\{D(t_d)t_d + \ln[1 + D(t_d)(n - t_d)]\}} \end{aligned} \quad (7)$$

where $D(t)$ is annual productivity decline rate in year t ; G is initial annual generation (kWh); N is power plant life (assumed to be 30 years in base case); C is total capital cost, that is, $c_d \cdot P$ (\$); c_o is unit annual O&M cost (¢/kWh); i is annual interest rate (assumed to be 7% in base case); I is annual inflation rate (assumed to be 3% in base case); c_{ofi} is fixed portion of the annual O&M cost at plant start-up divided by initial annual generation (¢/kWh); c_{ov} is variable portion of the annual O&M cost divided by annual generation (¢/kWh); N_{wi} is number of initial production wells; and C_{wi} is drilling cost per initial production well (assumed to be \$2 million in the base case).

Capital cost includes exploration cost, power plant cost, gathering and injection system cost, and cost of capital. Annual O&M cost includes personnel cost, general and administrative cost, insurance cost, supplies/consumables/engineering and laboratory services cost, wellfield maintenance cost, generator and turbine maintenance cost, and other equipment and maintenance cost.

The variable portion of the annual O&M cost represents costs that vary with the level of generation, such as, costs of supplies, consumables, etc., which remain proportional to generation; this cost divided by annual generation gives c_{ov} . The fixed portion of the annual O&M cost represents costs that are independent of the generation level; these include costs of personnel, administration, insurance,



Geothermal Power Economics. Figure 1

Schematic generation and make-up well drilling histories of a project [1]

wellfield maintenance, generator and turbine maintenance, other equipment maintenance, etc., which may not decline in response to any decline in generation. This fixed annual cost divided by annual generation gives c_{of} . For the purposes of this entry, 20% of the annual O&M cost was assumed to vary with generation at plant start-up; however, results are found to be relatively insensitive to the fraction of O&M cost that is variable. As generation declines, c_{ov} remains constant but c_{of} increases from its initial value of c_{ofi} . A typical plant capacity factor of 90% was assumed in estimating annual generation. In Eq. 7, the total capital cost (C) is assumed to be amortized over the plant life of n years at an interest rate i (annual compounding). The calculated power costs in future years are discounted for inflation to arrive at a levelized power cost in present dollars (\bar{c}).

Sensitivity of Levelized Power Cost

It should be noted that if there were no economy of scale in capital and O&M costs (i.e., a capital cost of \$2,184/kW and an O&M cost of 1.79 ¢/kWh, as in the base case) and if productivity decline rate were insensitive to installed capacity (remaining at 5% initial annual harmonic rate as in the base case), levelized power cost from Eq. 7 would be 3.6 ¢/kWh irrespective of plant capacity. Table 1 lists all parameters for the range of development scenarios analyzed, assuming the economy of scale in capital and O&M costs as well as the sensitivity of productivity decline to plant capacity.

Figure 2 shows the calculated power cost in ¢/kWh for various levels of installed plant capacity as a function of t_d (i.e., the number of years of make-up well drilling undertaken to maintain plant capacity). This figure takes into account the economy of scale as reflected in Eqs. 1 and 2, as well as acceleration in well productivity decline, as given by Eq. 5, with increased installed capacity. Figure 2 indicates that power cost declines with the number of years of make-up well drilling, the decline rate being steeper for a higher plant capacity. Figure 2 also indicates that if make-up well drilling is discontinued too early (prior to about 10 years), power cost would be higher for a larger plant. This figure also shows that for any plant capacity, a relatively minor reduction in power cost is achieved by continuing make-up well drilling after this period, and continuing make-up well drilling beyond about 20 years may actually increase power cost. Therefore, there is little reason to continue make-up well drilling beyond about 20 years unless the power sales contract imposes significant penalties for any shortfall in plant capacity.

Figure 3 shows the minimum achievable power cost for various plant capacities as read from Fig. 2. This figure shows that the minimum achievable power cost is rather insensitive to plant capacity; it varies from 3.7 ¢/kWh for a 10 MW plant to 3.4 ¢/kWh for a 150 MW plant, a 7.6% decline in power cost for a 1,400% increase in power capacity. Irrespective of the plant capacity and the number of years of make-up well drilling, power cost as of 2005 could not be lowered

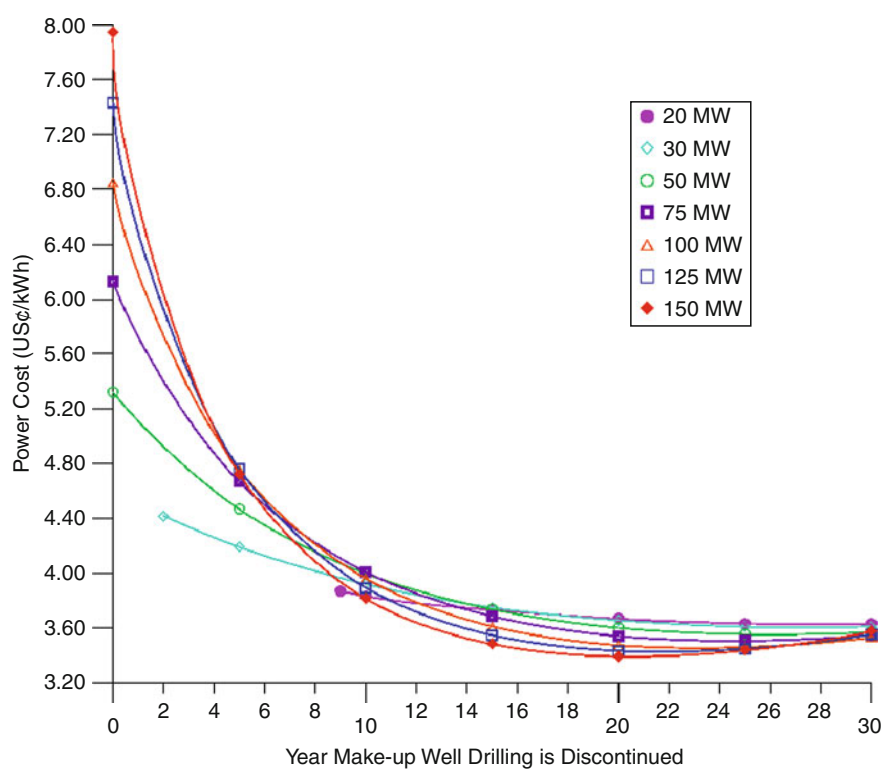
Geothermal Power Economics. Table 1 Development scenarios analyzed

Plant capacity (MW) ^c	Capital cost per kW	Total capital cost (million \$)	O&M cost (¢/kWh) ^b	Initial harmonic decline rate (%)	No. of initial production wells ^a	Years before make-up well drilling is required (t_c)
5	2,500	12.5	2.0	0.2	2	>30
10	2,463	24.6	1.98	0.6	3	>30
20	2,390	47.8	1.93	1.5	5	9
30	2,319	69.6	1.88	2.6	7	2
50	2,184	109.2	1.79	5.0	11	0
75	2,025	152.0	1.68	8.3	17	0
100	1,880	188.0	1.58	11.8	22	0
125	1,744	218.0	1.48	15.4	28	0
150	1,618	242.7	1.39	19.2	33	0

^a5 MW per well/minimum of one standby well/minimum of 10% excess capacity

^b80% of O&M cost varies with capacity

^cPlant capacity factor=0.9



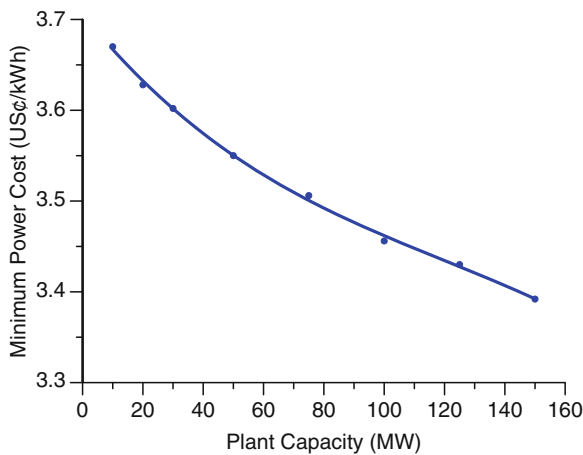
Geothermal Power Economics. Figure 2

Levelized power cost versus the year make-up well drilling is discontinued [1]

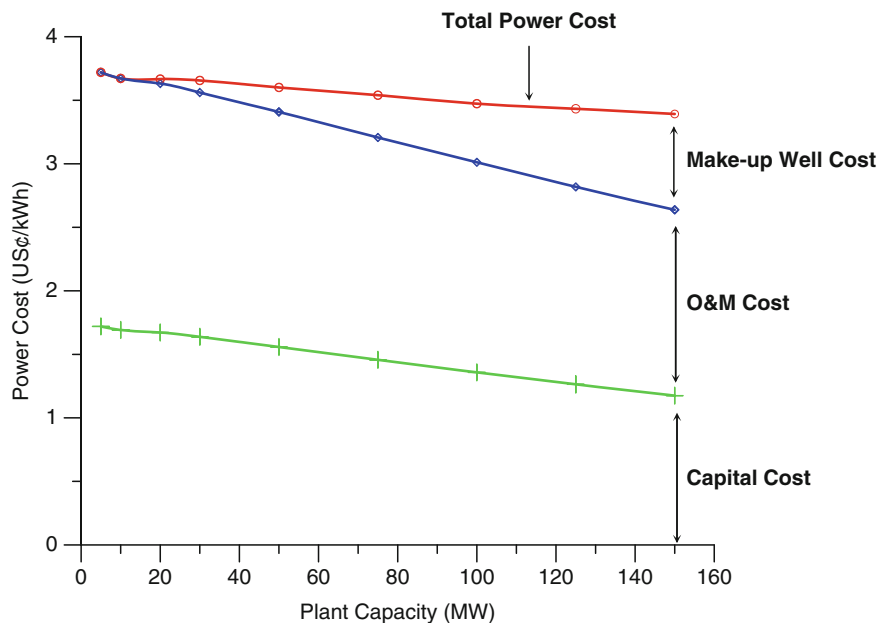
significantly below 3.4 ¢/kWh. Figure 4 shows the three components of power cost (capital, O&M, and make-up well drilling) as functions of plant capacity assuming make-up well drilling to be discontinued after 20 years. This figure shows that the capital cost component is approximately equal to the O&M cost

component for all plant capacities while the make-up well drilling component assumes greater significance with increasing plant capacity (except for very small capacities). Furthermore, the sum of O&M and make-up well drilling components constitutes the major part of power cost. Capital expenditure is incurred in the first few years of a project, when site-specific knowledge of the resource is still limited; therefore, adequate optimization of capital investment can be a challenge. After plant start-up, little can be done to reduce the capital cost component of power cost, except perhaps refinancing the debt should the interest rate decline. On the other hand, O&M and make-up well drilling costs, being incurred gradually as production continues, should reduce with time due to the “learning curve” effect. As more understanding of the resource characteristics and reservoir performance is gained with operation, O&M and make-up well drilling costs can be reduced, lowering power cost.

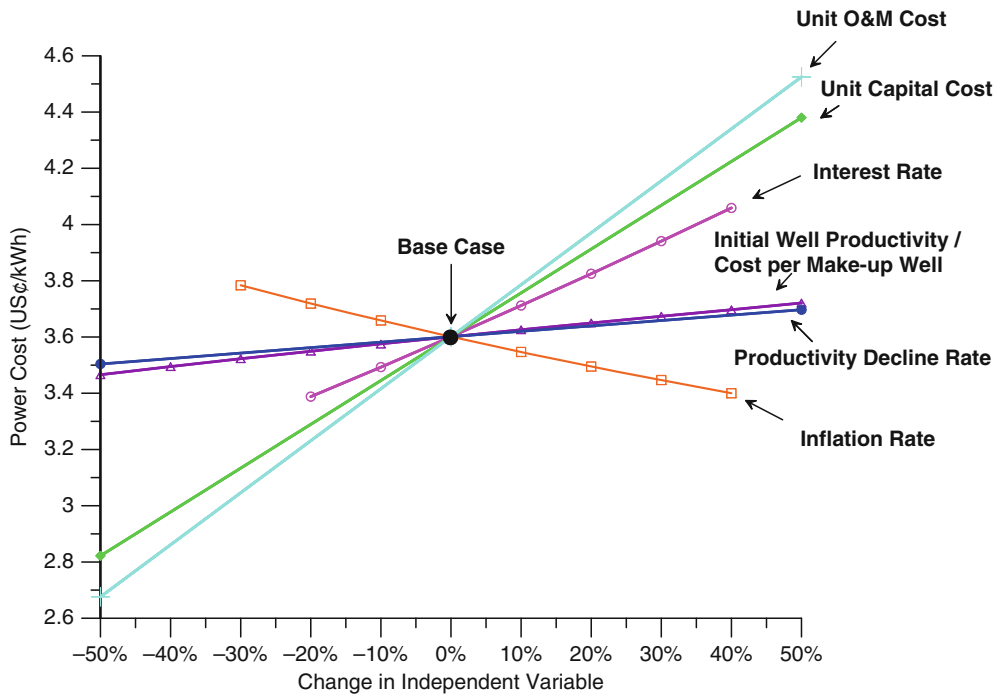
Figure 5 is a plot of power cost versus percent deviation in the values of the various independent variables from their base case (50 MW) values. In this figure, a steeper curve through the base case point implies a higher sensitivity of power cost to the variable



Geothermal Power Economics. Figure 3
Minimum levelized power cost versus plant capacity [1]



Geothermal Power Economics. Figure 4
Components of levelized power cost versus plant capacity (assuming 20 years of make-up well drilling) [1]



Geothermal Power Economics. Figure 5

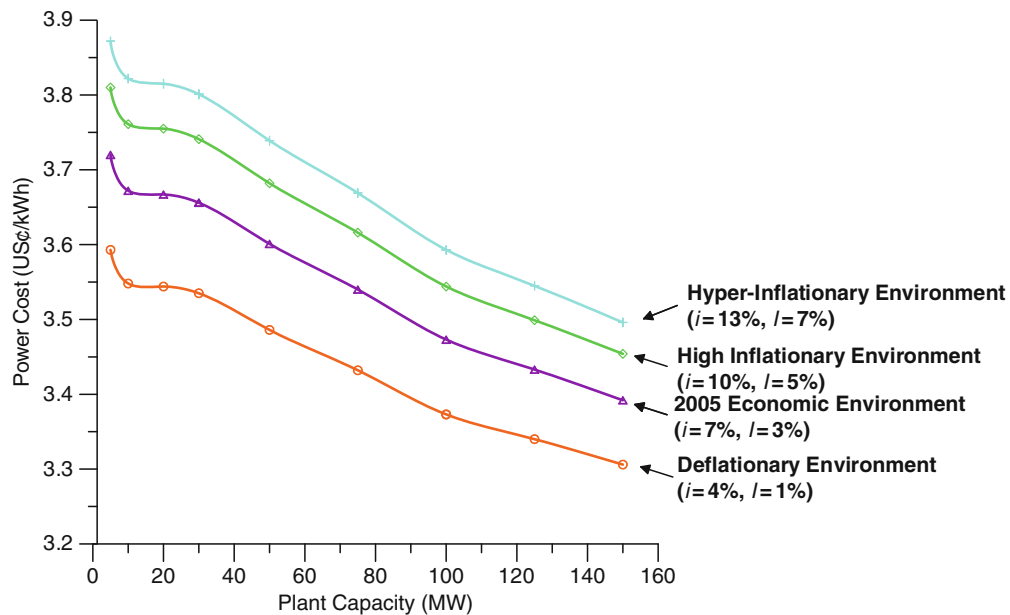
Sensitivity of base case power cost to changes in independent variables [1]

represented by the curve. Figure 5 shows that unit O&M cost and unit capital cost have the highest impact on power cost; these two variables are also subject to economy of scale. On the other hand, power cost is relatively insensitive to resource-related variables (such as well productivity, drilling cost per well, and productivity decline rate). Figure 5 indicates a levelized power cost of 3.6 ¢/kWh as of 2005 for a 50 MW plant. However, it should be noted that the author's experience as of 2005 indicated that the estimates of capital cost in the USA as of 2005 based on Ref. [2] was somewhat low. For the base case, the capital cost in the USA as of 2005 was as much as 30% higher than \$2,184/kW. Therefore, Figure 5 shows that the levelized power cost for a 50 MW plant in the USA as of 2005 was as high as 4.1 ¢/kWh; in 2010 it is as high as 8 ¢/kWh.

Interestingly, power cost is only modestly sensitive to macroeconomic variables (interest and inflation rates), because interest and inflation rates affect power cost by about the same magnitude but in opposite directions (Fig. 5). Figure 6 shows power cost versus plant capacity for several diverse microeconomic situations:

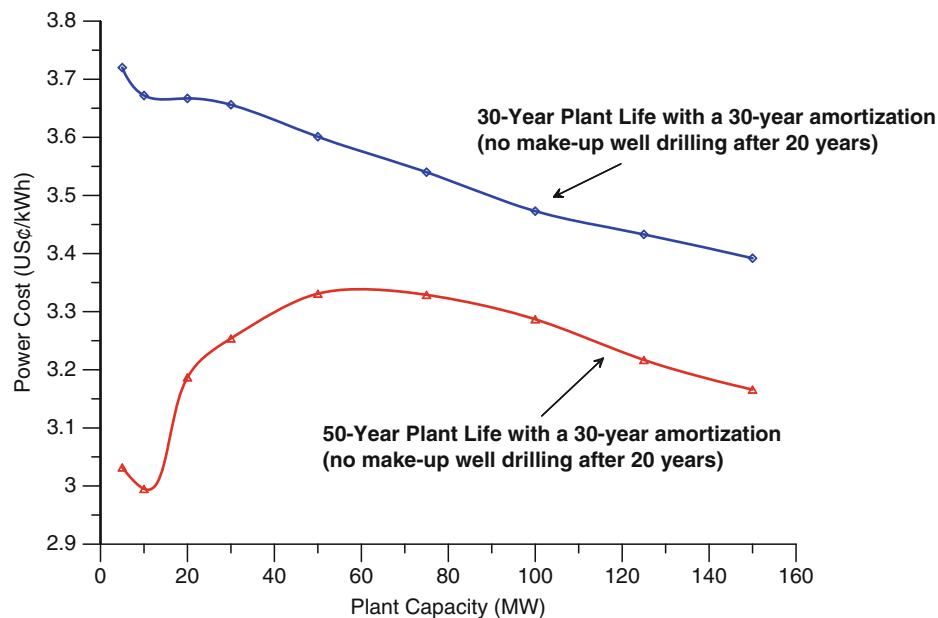
(1) a hyperinflationary environment, (2) a high inflationary environment, (3) the economic environment in the USA as of 2005, and (4) a deflationary environment; appropriate interest rates (i) and inflation rates (I) assumed for the various cases are shown on the figure. Figure 6 implies that, in relative terms, the sensitivity of power cost to the macroeconomic climate is not significant. For example, the variation in power cost over the capacity range of 5–150 MW is of similar magnitude as the variation in power cost in the base case over the extreme range of macroeconomic climates considered.

History of operation of geothermal power plants in Italy, New Zealand, El Salvador, Mexico, Japan, and USA, where some plants have already operated for more than 30 years, indicates that it is possible to continue operating a geothermal plant beyond its typical amortization period of 20–30 years. Can power cost be reduced if a geothermal plant were amortized for 30 years but operated for a longer period? Figure 7 compares power cost versus plant capacity as shown before (for 30 years' operation) and as calculated for



Geothermal Power Economics. Figure 6

Levelized power cost versus plant capacity under various macroeconomic conditions (for 20 years of make-up well drilling) [1]

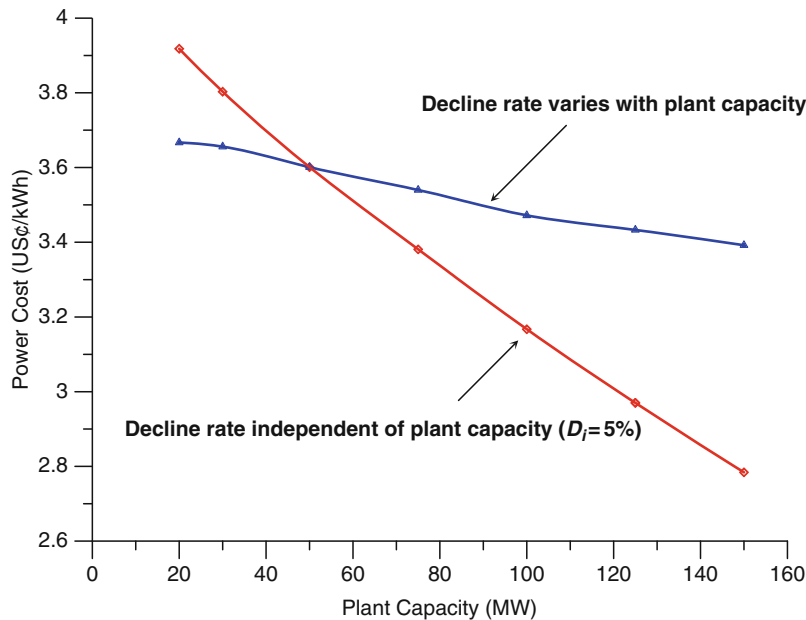


Geothermal Power Economics. Figure 7

Effect of plant life on levelized power cost (20 years of make-up well drilling) [1]

a 50-year operating period, the initial capital cost still being amortized over 30 years. Figure 7 shows that for smaller plants, cost may be reduced significantly, by as

much as 20% for plants of 10 MW or smaller capacity. For plants larger than about 50 MW, this reduction in power cost is not significant, particularly considering



Geothermal Power Economics. Figure 8

Levelized power cost versus plant capacity (for 20 years of make-up well drilling) [1]

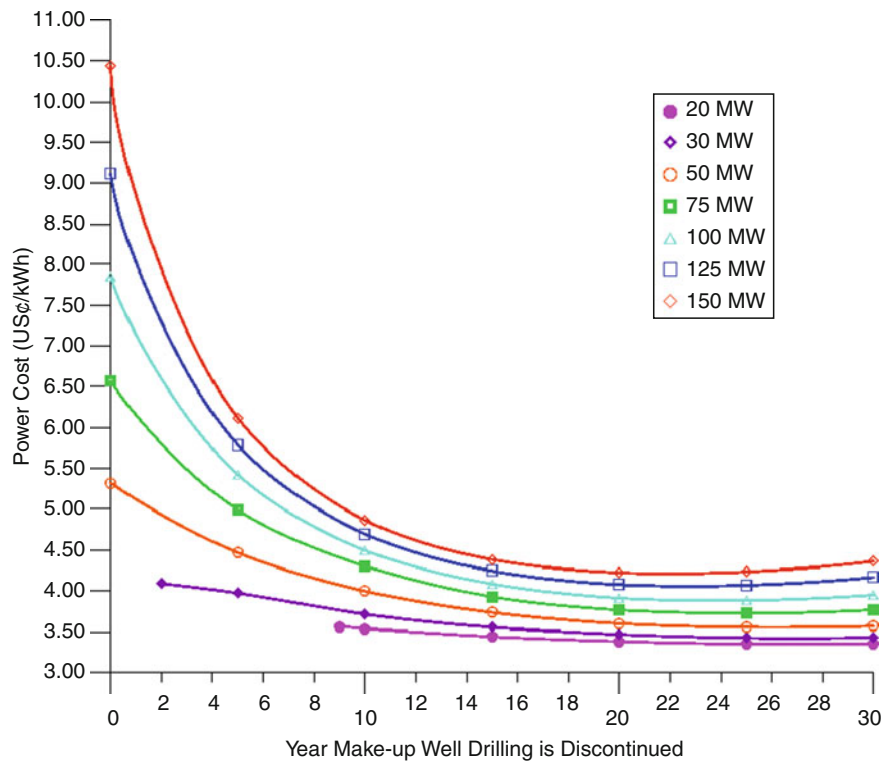
the additional risk of operating an aging power plant and pipelines, and possibly deteriorating wells.

The above analysis takes into account the usual acceleration in well productivity decline due to increases in plant capacity. How would the results change in the unusual case of well productivity being insensitive to installed plant capacity? Figure 8 compares levelized power cost as a function of plant capacity, as calculated before, with the case of a constant initial annual harmonic decline rate of 5% irrespective of capacity. Figure 8 shows that if productivity decline rate were insensitive to plant capacity, power cost would decline with plant capacity much more rapidly than in the usual case, the minimum power cost being only 2.8 ¢/kWh (for a 150 MW plant). However, a stand-alone project of a capacity larger than 100 MW is a rarity in the geothermal industry. The existing fields with a generation level greater than 100 MW typically rely on multiple, independent units of up to 100 MW each; as such, the economy of scale enjoyed by these projects would amount to that for a capacity of 100 MW or less. Therefore, if well productivity were insensitive to plant capacity, a power cost of less than 3.2 ¢/kWh (estimated for a 100 MW plant) as of 2005 was unlikely to be realized.

Finally, how would the results change if economy of scale in capital and O&M costs were negligible? One such conceivable situation could be the installation of multiple, modular and infrastructurally independent power plants in the same field. Figure 9 presents power cost versus the number of years of make-up well drilling for various plant capacities ignoring economy of scale. The results in this figure assume that unit capital and O&M costs remain the same as in the base case irrespective of installed capacity, but productivity decline still increases with installed capacity as given by Eq. 5. Figure 9 indicates that if economy of scale were negligible, power price would increase with installed capacity no matter how long one continues make-up well drilling, and power price would be consistently higher than in the usual case with economy of scale. The minimum achievable power cost in this case is still on the order of 3.4 ¢/kWh as of 2005 (estimated for a 20 MW plant).

Concluding Remarks

1. Power cost is sharply reduced by maintaining full generation capacity, by drilling make-up wells, for at least the first 10 years or so following plant



Geothermal Power Economics. Figure 9

Levelized power cost versus the year make-up well drilling is discontinued ("no economy of scale" case) [1]

- start-up; continuing make-up well drilling beyond 20 years does not reduce power cost significantly.
- The minimum achievable power cost is insensitive to plant capacity; as of year 2005, it was on the order of 3.4 ¢/kWh. There are significant opportunities to reduce power cost as site-specific experience is gained in resource management and power plant operation throughout the project life.
- The levelized cost of power from a 50 MW plant as of 2005 was in the range of 3.6–4.1 ¢/kWh; it is approximately twice as of 2010.
- Power cost is most sensitive to unit O&M cost followed by unit capital cost, interest rate, and inflation rate in the decreasing order of sensitivity; it is relatively insensitive to well productivity, drilling cost per well, well productivity decline rate, and the macroeconomic climate.
- Operating small power plants (10 MW or less capacity) beyond their typical amortization period of 30 years can significantly reduce power cost.
- The minimum achievable power cost does not decline significantly with increasing plant capacity except in the unlikely situation of well productivity decline being insensitive to capacity, when it was as low as 3.2 ¢/kWh as of 2005. In the unusual situation of an absence of economy of scale, power cost increases with plant capacity, the minimum achievable level being 3.4 ¢/kWh. In the very unlikely situation of both well productivity decline as well as unit capital and O&M costs being insensitive to plant capacity, the minimum achievable power cost would be on the order of 3.6 ¢/kWh in 2005 dollars.

Future Directions

This entry analyzes the levelized cost of geothermal power as of 2005. Between 2005 and 2010, this cost has escalated in spurts; today levelized cost is nearly double that in 2005. However, at this time the cost escalation is relatively slow.

While leveled cost of geothermal power has increased over the last 5 years, so has the price available for geothermal power in the USA by a similar ratio. The recent price increases have been driven by a host of new incentives for geothermal power introduced in the USA: (1) the Renewable Portfolio Standard (“RPS”), which requires an utility to derive a minimum fraction (0–33% depending on the State) of its power from renewable resources; (2) a Production Tax Credit (“PTC”) of about 2 ¢/kWh for geothermal power; (3) the option to receive an Investment Tax Credit (30% of the capital investment) at the onset of power generation in lieu of PTC; (4) a renewable energy credit (“REC”) in the USA or carbon credit worldwide; Geothermal Loan Guaranty, etc. Therefore, it is reasonable to expect that the price of geothermal power will continue to keep pace with, or most likely increase relative to, leveled cost of geothermal power for the near future.

It should be noted that this entry considers the economics of power from conventional geothermal systems, which are naturally occurring subsurface porous or fractured systems that can be tapped for production by drilling wells. However, in the last decade, considerable hopes have been raised of tapping geothermal energy from enhanced geothermal systems (“EGS”) [5]. These are hot subsurface systems with porosity or fracture capacity too low to allow commercial production but can be enhanced by pervasive hydraulic fracturing to enable significant fluid injection and production. In an EGS project, heat is recovered from the artificial reservoir by injecting cool water through a set of wells while producing heated water from another set of wells. Such systems have not yet proven commercial, but research and development toward commercial tapping of EGS systems continue. Even in countries where conventional geothermal systems do not exist, EGS developments would be the theoretically possible, because anywhere on earth adequately hot rock bodies can be reached by drilling wells deep enough and creating an artificial reservoir by hydraulic fracturing of rock. Sanyal et al. [6] has presented an analysis of the economics of a prototype EGS project. However, until EGS power proves commercial, its economics would have significant uncertainties.

Bibliography

1. Sanyal SK (2005) Levelized cost of geothermal power – how sensitive it is? *Trans Geotherm Res Council* 29:459
2. Entingh DJ, McVeigh JF (2003) Historical improvements in geothermal power systems costs. *Trans Geotherm Res Council*, 533
3. Sanyal SK, Menzies AJ, Brown PJ, Eneedy KL, Eneedy S (1989) A systematic approach to decline curve analysis for the geysers steam field, California. *Trans Geotherm Res Council*, Santa Rosa, p 415
4. Sanyal SK, Robertson-Tait A, Klein CW, Butler SJ, Lovekin JW, Brown PJ, Sudarman S, Sulaiman S (2000) Assessment of steam supply for the expansion of generation capacity from 140 to 200 MW, Kamojang Geothermal Field, West Java Indonesia. In: *Trans World Geotherm Congress*, Beppu and Morioka, Japan, p 2195
5. MIT (2006) The future of geothermal energy – impact of enhanced geothermal systems (EGS) on the United States in the 21st century. An assessment by an MIT – led interdisciplinary panel, Massachusetts Institute of Technology, Cambridge
6. Sanyal SK, Morrow JW, Butler SJ, Robertson-Tait A (2007) Is EGS commercially feasible? *Trans Geotherm Res Council* 31:2007

Geothermal Power Stations, Introduction to

LUCIEN Y. BRONICKI

Ormat Technologies, Inc., Reno, NV, USA

Geothermal energy, contrary to solar and wind, does not depend on weather and is the only one which can supply base-load power. Although not evenly distributed geographically, geothermal energy potential is very important, particularly if the R&D of such advanced systems will be actively pursued.

This section covers the nature of geothermal energy resources, their utilization, conversion technologies as well as its future development (► [Geothermal Energy, Nature, Use, and Expectations](#)). The section also highlights the greatest challenge in geothermal development, namely, the geothermal resource. Power conversion is the least uncertain part of a geothermal project, as it consists of a straightforward engineering design with work executed by experienced manufacturers, engineering firms, and contractors.

The risks and challenges are related to exploration, drilling, and managing the resource. Optimization

depends on the choice of adaptation of the power station configuration to the resources available.

When considering different types of resources and uses of geothermal energy, it is important to differentiate between geothermal or ground source heat pumps (GSHP) which utilize the natural insulation of the Earth for heating and cooling, as compared to hydrothermal resources, which are natural flows of geothermal-heated waters associated with underground heat sources or hot rocks.

- The comparative advantages of geothermal energy use include low emissions, high capacity factors, sustainability, and minimal land footprints.
- Hydrothermal sources can be utilized directly for space heating and other direct uses. This source can also be used for power generation if fluids of sufficient temperatures are available at commercial depths.

The distribution of geothermal resources is irregular due to unequal distribution of volcanoes, hot springs, and heat manifestations at specific locations over the Earth's surface (► [Geothermal Energy, Geology and Hydrology of](#)). Geothermal resources are a reflection of the underlying global, local geological and hydrological framework. The most thermally rich resources tend to concentrate in environments with abundant volcanic activity and tend to be controlled by plate tectonic processes or spreading centers evident as, volcanic chains associated with subduction zones and hot spots. The local geological characteristics that favor useful resources include relatively shallow resource depths to high permeability in the rocks surrounding the resource, and adequate resource fluids.

Exploration starts with the analysis of available geological information to identify the potential target. Once the target is identified, geochemical studies and core drilling are undertaken. These studies are complemented or sometimes preceded by geophysical surveys including aeromagnetic or resistivity studies and remote infrared and hyperspectral techniques.

Hydrothermal systems have different chemical properties (► [Hydrothermal Systems, Geochemistry of](#)). The source of heat is usually a magma chamber a few kilometers below the surface. Less frequently, the

source of heat is a crustal site. Fluid origin is meteoric, that is, rainwater which infiltrates the ground to depths of a few kilometers. The permeability and degree of fracturing of this cap rock varies from site-to-site according to the intensity and abundance of the surface hydrothermal systems manifestations (boiling springs, steam vents, hot ponds, and geysers).

In the early stages of a hydrothermal system exploration, when there is only surface evidence, the aim of a geochemical survey is the generation of a model that evaluates the temperature and chemical conditions of the fluid at depth.

Drilling is the process for extracting geothermal energy resources for energy production utilization (► [Geothermal Resources, Drilling for](#)). Shallow or intermediate-depth wells may be drilled for the purposes of space heating or direct heat uses; more substantial drilling activities are needed to drill for hotter resources designed for power generation.

Geothermal drilling is a niche within the larger drilling services industry which focuses primarily on oil, gas, and minerals. In particular, deep drilling, required in most exploration programs for geothermal power generation projects, will likely utilize big rigs typical in oil and gas extraction.

There are several aspects unique to geothermal drilling. Mainly, geothermal formations, by nature, involve elevated temperatures which are usually significantly higher than those experienced when drilling for oil and gas. The rock that hosts these formations are typically harder (granite, granodiorite, quartzite, basalt, volcanic tuff), more abrasive, highly fractured, and under-pressured. Caustic elements may be present that can cause corrosion and scaling in the wellbore.

These unique characteristics present challenges in dealing with geothermal wells which, unlike oil and gas wells, do not produce economically until utilized through electric generation or direct uses. For power production, geothermal wells must be of a larger diameter than oil and gas wells to produce necessary flow rates for commercial production. Depths of geothermal wells vary according to location. Some resources are shallow (<1,000 m) and others deep (2,500 m to over 3,000 m).

Reservoir Engineering is the comprehensive integration of all available surface and underground information

regarding geology, geophysics, geochemistry, well drilling-testing, exploitation data, information concerning the geothermal developer and objectives of a geothermal development: market targets, costs, and finance becoming the most powerful tools to evaluate the feasibility (► [Reservoir Engineering in Geothermal Fields](#)). As in any scientific or engineering activity, results derived from reservoir engineering depend on the quantity and quality of collected information and the as well as the assimilated handling and in depth comprehension of the collected information. Reservoir engineering is not limited to the final numerical tool, but also includes acquisition information which allows prediction of the impacts on a geothermal resource 20–30 years into the future.

Maintaining the sustainability of a geothermal field through operation requires Monitoring (► [Geothermal Field and Reservoir Monitoring](#)). Using techniques such as down-hole monitoring and surface monitoring, the impact of production on the long-term sustainability of a geothermal field can be closely evaluated to ensure that the resource is not prematurely cooling and that any cooling is minimal.

The heat stored in hot dry rock is not accessible via conventional geothermal technology. New methods are being developed and tested to access the huge potential of this type of resources (► [Engineered Geothermal Systems, Development and Sustainability of](#)). Commercialization of this technology could unlock many thousands of megawatts of power. For example, the estimate of the Technical Potential for EGS in the USA is estimated at 100 GWe, which is 30 times the total current installed US geothermal capacity from all energy sources. There has been some success, but no actual production from EGS reservoirs as of the end of 2010. Once commercialized, EGS needs to be proved sustainable.

As with any other geothermal energy source, EGS development involves some impact on the environment (► [Geothermal Resources, Environmental Aspects of](#)). Geothermal resources are environmentally important as natural thermal features. Typically, the most significant environmental impacts are associated with the exploitation of high-temperature liquid-dominated geothermal systems for electric power generation; however, the majority of these impacts can be avoided or minimized with appropriate

techniques. However, as geothermal energy generally offsets use of fossil fuels, the use of geothermal resources are more likely to improve air quality and overall water quality.

The current utilization of geothermal resources worldwide includes direct use of heat and power generation (► [Geothermal Energy Utilization](#)). There is a long history of using geothermal heat since the Roman times. The development of geothermal usage began early in the twentieth century in industrial countries with abundant geothermal activity. The list of these countries includes Italy, USA, Japan, New Zealand and Iceland. In addition, developing countries such as the Philippines, Indonesia, Central America and Kenya are also listed as geothermal users.

Applications, such as space heating, agriculture, and other processes, require heat that may otherwise be provided using a fossil fuel (► [Geothermal Resources Worldwide, Direct Heat Utilization of](#)). Recent developments in using geothermal sources in large-scale projects include district heating, greenhouse complexes, and major industrial uses. Heat exchangers are also becoming more efficient and better adapted to geothermal projects. This allows the use of lower-temperature water and highly saline fluids. Heat pumps utilizing very low-temperature fluids have extended geothermal developments into traditionally non-geothermal countries such as Canada, France, and Switzerland, as well as areas of the USA.

With the end of 2010, global use of direct geothermal utilization added to approximately 50.5 GW of thermal energy, in 78 countries, displacing over 121 TWh/year of energy consumption.

The techniques used for the conversion of geothermal fluid heat content into mechanical power are similar to those used in fossil-fueled power plants (► [Geothermal Power Conversion Technology](#)). Power conversion is the most predictable part of a geothermal project, as it consists of a straightforward engineering design with work executed by experienced manufacturers, engineering firms, and contractors.

The risks and challenges are related to exploration, drilling, and managing the resource. Optimization depends on the adaptation of the power station configuration to the available resources.

Today, 10,000 MW of geothermal power plants are in operation and a majority of them use steam turbines that operate on dry steam or steam produced by single or double flash with about 1,000 MW using Organic Rankine Cycles (ORC) or geothermal combined cycles. However, to widen the range of resources suitable for power generation beyond dry steam and flashed steam plants, ORC cycles have been implemented in the last 30 years, and will probably continue to grow as a common technology driving future development of geothermal resources.

Operational experience confirms the advantages of ORC power stations, not only for low-enthalpy water-dominated resources, but also for certain high-enthalpy sources where the brine is aggressive or the fluid contains a high percentage of non-condensable gas. The higher installation cost of these systems is often justified by environmental and long-term resource management considerations.

In geothermal systems, it is possible to estimate the commercial, sustainable, and renewable capacities of a geothermal system (► [Geothermal Power Capacity, Sustainability and Renewability of](#)). Sustainability is defined as the ability to economically maintain an installed power capacity over the amortized life of a power plant. This is done by taking practical steps, such as drilling “make-up” wells as required to compensate for resource degradation. Renewability is defined as the ability to maintain an installed power capacity indefinitely without encountering any resource degradation. Typically, the renewable power capacity at a geothermal site is generally too small for commercial development of electrical power capacity, but may be adequate for district heating or other direct uses of the geothermal energy.

The cost in producing geothermal resources for electric generation is important (► [Geothermal Power Economics](#)). In particular, the levelized cost of power is the applicable measurement for the cost of geothermal energy. Unlike fossil fuel power plants, most of the capital costs are incurred upfront in the development of the resource. Power cost is an objective criterion that favors the geothermal solutions compared to other alternative energy sources. However, the costs are heavily tied to the resource and the need for make-up well drilling to maintain full generation capacity.

Geothermal Resources Worldwide, Direct Heat Utilization of

JOHN W. LUND

Geo-Heat Center, Oregon Institute of Technology,
Klamath Falls, OR, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Direct-Use Temperature Requirements
Equipment
Economic Considerations
Energy Savings
Future Directions
Bibliography

Glossary

Agribusiness applications In the geothermal context, they are the heating of greenhouses and open ground for various crops, aquaculture ponds and raceways heating for various aquatic species, and the heating of animal pens and houses in an effort to increase production and shortening the growing cycle.

Balneology The science of healing qualities of baths, especially with natural mineral waters and the therapeutic use of natural warm or mineral waters.

District heating Heating of more than one building from a central heating plant with the heated fluid provided through a central distribution systems of pipes.

Heat exchanger A device for transferring heat from one fluid to another. The fluids are usually separated by conducting walls of metal or plastic.

Heat pump A device which, by the consumption of work or heat, effects the transport of heat between a lower temperature to a higher temperature source. The useful output is heat in conventional usage. The reverse process is called a refrigerator used for the removal of heat.

Joule (J) The SI unit for all forms of energy or work. It is equal to 1 W-s or 0.239 cal.

Spa A resort using mineral water for bathing, soaking, and drinking along with covering portions of the body with mineral muds for therapeutic purposes. Diet, exercise, and rest can also be part of the spa treatment plan.

Watt (W) A unit of power or energy produced over time, equivalent to 1 J/s, or 0.001341 horse power (hp).

Definition of the Subject

Direct or non-electric utilization of geothermal energy refers to the immediate use of the heat energy rather than to its conversion to some other form such as electrical energy. The primary forms of direct-use include heating swimming pools and baths, and for balneology (therapeutic use), space heating and cooling including district heating, agriculture (mainly greenhouse heating, crop drying, and some animal husbandry), aquaculture (mainly fish pond and raceway heating), providing heat for industrial processes, and heat pumps (for both heating and cooling). In general, the geothermal fluid temperatures required for direct heat use are lower than those for economic electric power generation, and as a result these resources are available in most countries.

Most direct-use applications use geothermal fluids in the low-to-moderate temperature range between 50°C and 150°C, and in general, the reservoir can be exploited by conventional water well drilling equipment. Low-temperature systems are also more widespread than high-temperature systems (above 150°C), so they are more likely to be located near potential users. In the USA, for example, of the 1,350 known or identified geothermal systems, 5% are above 150°C, and 85% are below 90°C [1]. In fact, almost every country in the world has some low-temperature systems, while only a few have accessible high-temperature systems.

Geothermal energy is a renewable energy since the tapped heat is continuously renovated by natural process of the Earth's interior, and the extracted geothermal fluids are replenished by natural recharge and by reinjection of the exhausted fluids, providing a sustainable development. Using geothermal minimizing the greenhouses gases and particulates that are produced from using fossil

fuels, and also provides energy independence since it is a domestic resource. The environmental impact of direct-use of geothermal energy is negligible as in most cases, once the heat is extracted from the fluid, the spent fluid is reinjected back into the ground, thus preventing the release of harmful gasses and particulates.

Introduction

Traditionally, direct use of geothermal energy has been on small scale by individuals. More recent developments involve large-scale projects, such as district heating (Iceland and France), greenhouse complexes (Hungary and Russia), or major industrial use (New Zealand, Iceland, and the USA). Heat exchangers are also becoming more efficient and better adapted to geothermal projects, allowing use of lower temperature water and highly saline fluids. Heat pumps utilizing very low-temperature fluids (5–30°C) have extended geothermal developments into traditionally non-geothermal countries such as Canada, France, Switzerland, and Sweden, as well as areas of the mid-western and eastern USA. Most equipments used in these projects are of standard, off-the-shelf design and need only slight modifications to handle geothermal fluids [2, 3].

Worldwide [4], the installed capacity of direct geothermal utilization is 50,583 MWt, and the energy use is 438,071 TJ/year (121,686 GWh/year), distributed among 78 countries; the leading countries are presented in Table 1. This amounts to saving an equivalent 45.2 million tons of fuel oil per year (TOE) if it replaces electricity. The distribution of the energy use among the various types is listed in Table 2 and shown in Fig. 1 for the worldwide installed capacity, and Fig. 2 for the annual energy use. For comparison, the installed capacity in the USA (2010) is 12,611 MWt, and the annual energy use is 56,552 TJ (15,709 GWh), saving 20.2 million barrels of oils (3.04 million TOE) [5]. Internationally, the largest energy uses are for geothermal heat pumps (GHP) (49%), and swimming, bathing, and balneology (25%); and similar, in the USA, the largest use is for geothermal heat pumps (84%). In comparison, Iceland's largest geothermal energy use is 72% for district heating 17,483 TJ/year (4,857 GWh/year) [6]. As can be seen from Tables 1 and 2, heat pumps have low load factors (USA), whereas industrial

Geothermal Resources Worldwide, Direct Heat Utilization of. Table 1 The leading direct-use countries (2010)

Country	Energy use		Power MWt	Capacity Factor	Main applications
	TJ/year	GWh/year			
China	75,348	20,932	8,898	0.27	Bathing/district heating
USA	56,552	15,710	12,611	0.14	GHP
Sweden	45,301	12,585	4,460	0.32	GHP
Turkey	36,886	10,247	2,084	0.56	District heating
Japan	25,698	7,139	2,100	0.39	Bathing (onsens)
Iceland	24,361	6,768	1,826	0.42	District heating
France	12,929	3,592	1,345	0.30	District heating
Germany	12,764	3,546	2,485	0.16	Bathing/district heating
The Netherlands	10,699	2,972	1,410	0.24	GHP
Italy	9,941	2,762	867	0.36	Spas/space heating
Hungary	9,767	2,713	655	0.47	Spas/greenhouses
New Zealand	9,552	2,654	393	0.77	Industrial uses
Canada	8,873	2,465	1,126	0.25	GHP
Switzerland	7,715	2,143	1,061	0.23	GHP

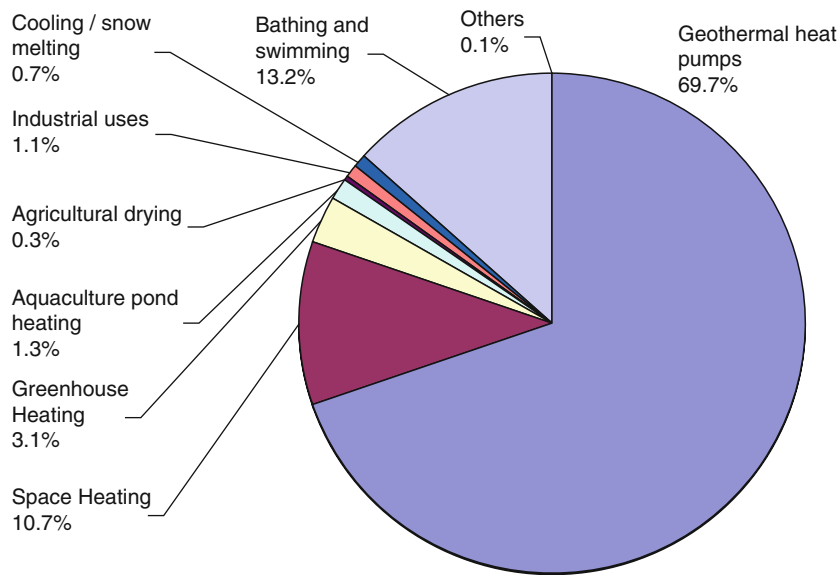
Geothermal Resources Worldwide, Direct Heat Utilization of. Table 2 Summary of geothermal direct-use by category (2010)

Category	Utilization		Capacity (MWt)	Capacity factor
	(TJ/year)	(GWh/year)		
Geothermal heat pumps	214,782	59,662	35,236	0.19
Space heating	62,984	17,496	5,391	0.37
Greenhouse heating	23,264	6,462	1,544	0.48
Aquaculture pond heating	11,521	3,200	653	0.56
Agricultural drying	1,662	462	127	0.42
Industrial uses	11,746	3,263	533	0.70
Bathing and swimming	109,032	30,287	6,689	0.52
Cooling/snow melting	2,126	591	368	0.18
Others	956	266	41	0.73
Total	438,071	121,686	50,583	0.27

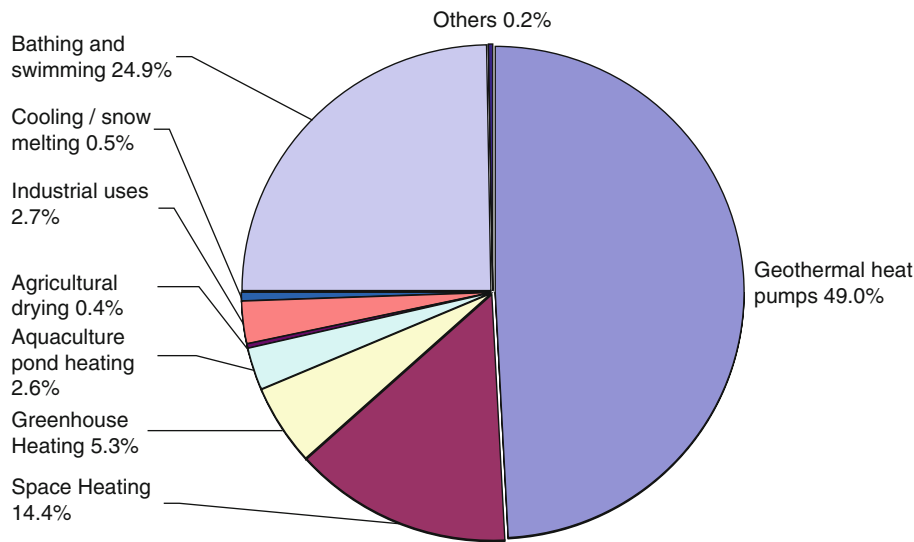
uses have high load factors (NZ) due to the more continuous annual use in industrial processing.

The Lindal diagram [7], named for Baldur Lindal, the Icelandic engineer who first proposed it, indicates

the temperature range suitable for various direct-use activities (Fig. 3). Figure 4 indicates some of the worldwide direct-use applications along with their possible temperature range of use. Typically, the agricultural



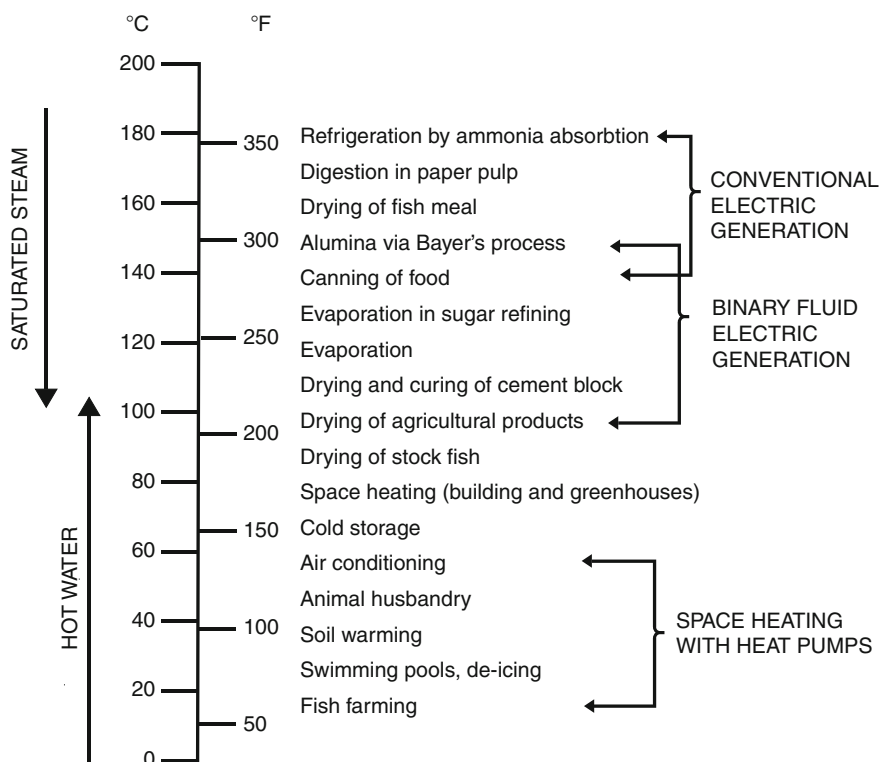
Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 1
Distribution of direct-use installed capacity (MWt) in the world (2010)



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 2
Distribution of direct-use annual energy use (TJ/year) in the world (2010)

and aquacultural uses require the lowest temperatures, with values from 25°C to 90°C. The amounts and types of chemicals, such as arsenic and dissolved gases such as boron, are a major problem with plants and animals; thus, heat exchangers are often necessary. Space heating

requires temperatures in the range of 50–100°C, with 40°C useful in some marginal cases and ground-source heat pumps extending the range down to 5°C. Cooling and industrial processing normally require temperatures over 100°C. The leading user of geothermal



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 3
Lindal diagram

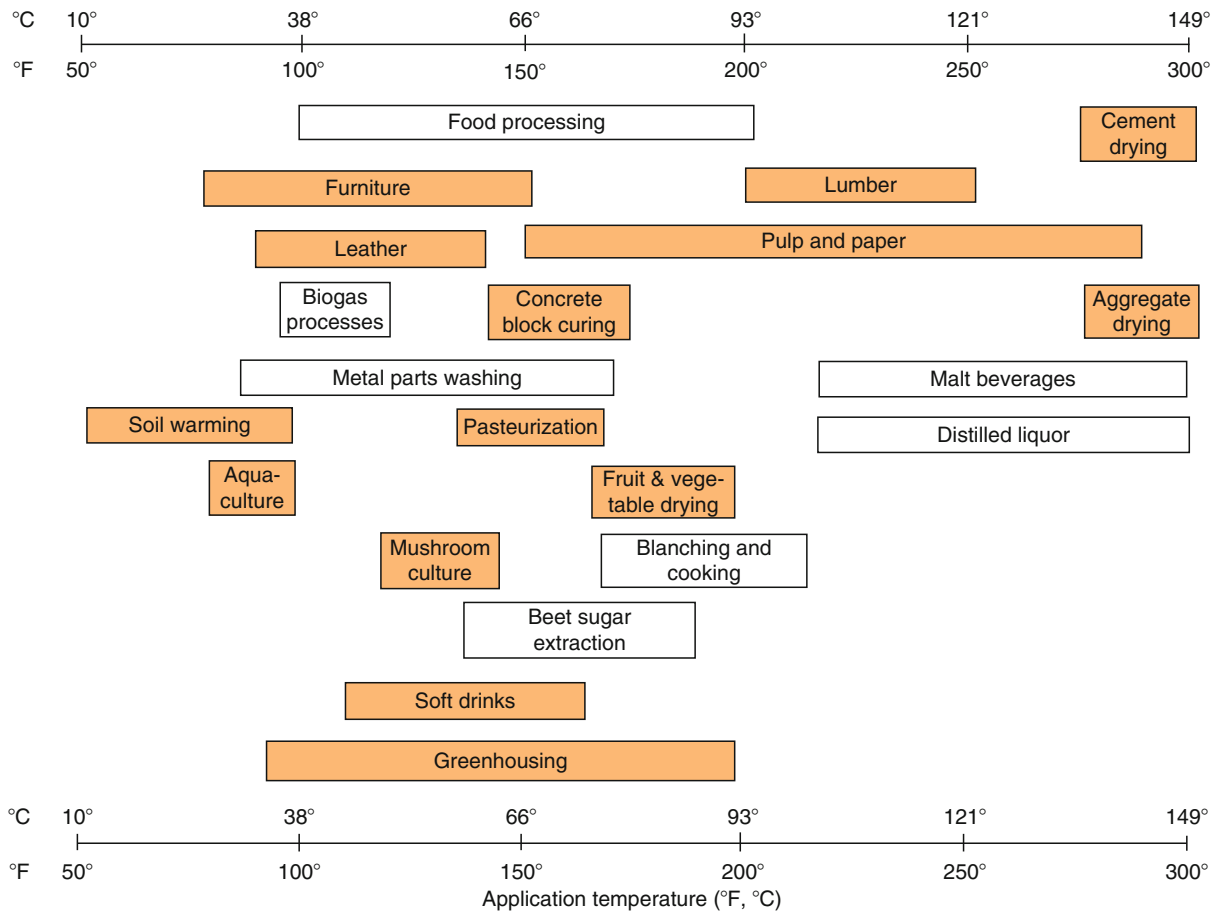
energy, in terms of market penetration, is Iceland, where more than 89% of the population enjoys geothermal heat in their homes from 30 municipal district heating services, and 54% of the country's total energy use is supplied by direct heat and electrical energy derived from geothermal resources [6].

Swimming, Bathing, and Balneology

People have used geothermal water and mineral waters for bathing and their health for many thousands of years. Balneology, the practice of using natural mineral water for the treatment and cure of disease, also has a long history. A spa originates at a location mainly due to the water from a spring or well. The water, with certain mineral constituents and often warm, give the spa certain unique characteristics that will attract customers. Associated with most spas is the use of muds (peoloids) which either are found at the site or are imported from special locations. Drinking and bathing in the water, and using

the muds are thought to give certain health benefits to the user. Swimming pools have desirable temperature at 27°C; however, this will vary from culture to culture by as much as 5°C. If the geothermal water is higher in temperature, then some sort of mixing or cooling by aeration or in a holding pond is required to lower the temperature, or it can first be used for space heating, and then cascaded into the pool. If the geothermal water is used directly in the pool, then a flow-through process is necessary to replace the "used" water on a regular basis. In many cases, the pool water must be treated with chlorine; thus, it is more economical to use a closed loop for the treated water and have the geothermal water provide heat through a heat exchanger [8].

Romans, Chinese, Ottomans, Japanese, and central Europeans have bathed in geothermal waters for centuries. Today, more than 2,200 hot springs resorts in Japan draw 100 million guests every year, and the "return-to-nature" movement in the USA has revitalized many hot spring resorts.



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 4
Examples of industrial applications of geothermal energy with the colored bars indicating those currently using geothermal energy in the world

The geothermal water at Xiaotangshan Sanitarium, northwest of Beijing, China, has been used for medical purposes for over 500 years. Today, the 50°C water is used to treat high blood pressure, rheumatism, skin disease, diseases of the nervous system, ulcers, and generally for recuperation after surgery. In Rotorua, New Zealand, at the center of the Taupo Volcanic Zone of North Island, the Queen Elizabeth Hospital was built during World War II for US servicemen and later became the national hospital for the treatment of rheumatic disease. The hospital has 200 beds, and outpatient service, and a cerebral palsy unit. Both acidic and basic geothermally heated mud baths treat rheumatic diseases.

In Beppu, on the southern island of Kyushu, Japan, the hot water and steam meet many needs: heating,

bathing, cooking, industrial operations, agriculture research, physical therapy, recreational bathing, and even a small zoo [9]. The waters are promoted for “digestive system troubles, nervous troubles, and skin troubles.” Many sick and crippled people come to Beppu for rehabilitation and physical therapy. There are also eight Jigokus (hot springs or geysers called “burning hells”) in town, showing various geothermal phenomena, used as tourist attractions.

In the former Czechoslovakia, the use of thermal waters has been traced back before the occupation of the Romans and has had a recorded use of almost 1,000 years. Today, there are 60 spa resorts located mainly in Slovakia, visited by 460,000 patients usually for an average of three weeks each. These spas have old

and well-established therapeutic traditions. Depending on the chemical composition of the mineral waters and spring gas, availability of peat and sulfurous mud, and climatic conditions, each sanatorium is designated for the treatment of specific diseases. The therapeutic successes of these spas are based on centuries of healing tradition (balneology), systematically supplemented by the latest discoveries of modern medical science [10].

Bathing and therapeutic sites in the USA include: Saratoga Springs, New York; Warm Springs, Georgia; Hot Springs, Virginia; White Sulfur Springs, West Virginia; Hot Spring, Arkansas; Thermopolis, Wyoming; and Calistoga, California. The original use of these sites was by Indians, where they bathed and recuperated from battle as neutral ground. There are over 115 major geothermal spas in the USA with an annual energy use of 1,500 TJ [8].

Figures for this use are difficult to collect and quantify. Almost every country has spas and resorts that have swimming pools (including balneology), but many allow the water to flow continuously, regardless of use. As a result, the actual usage and capacity figures may be high. Undeveloped natural hot springs have not been included in the data. A total of 67 countries have reported bathing and swimming pool use, amounting to a worldwide installed capacity of 6,689 MWt and energy used of 109,032 TJ/year (30,287 GWh/year) based on data from country update papers from the World Geothermal Congress 2010 (WGC2010) in Bali, Indonesia [4].

Space Conditioning

Space conditioning includes both heating and cooling. Space heating with geothermal energy has widespread application, especially on an individual basis. Buildings heated from individual wells are popular in Klamath Falls, Oregon; Reno, Nevada, USA; and Taupo and Rotorua, New Zealand. Absorption space cooling with geothermal energy has not been popular because of the high-temperature requirements and low efficiency. However, newer units recently placed on the market report to use temperatures below 100°C efficiently. Geothermal heat pumps (groundwater and ground-coupled) have become popular in the USA, Canada, China, and Europe, used for both heating and cooling.

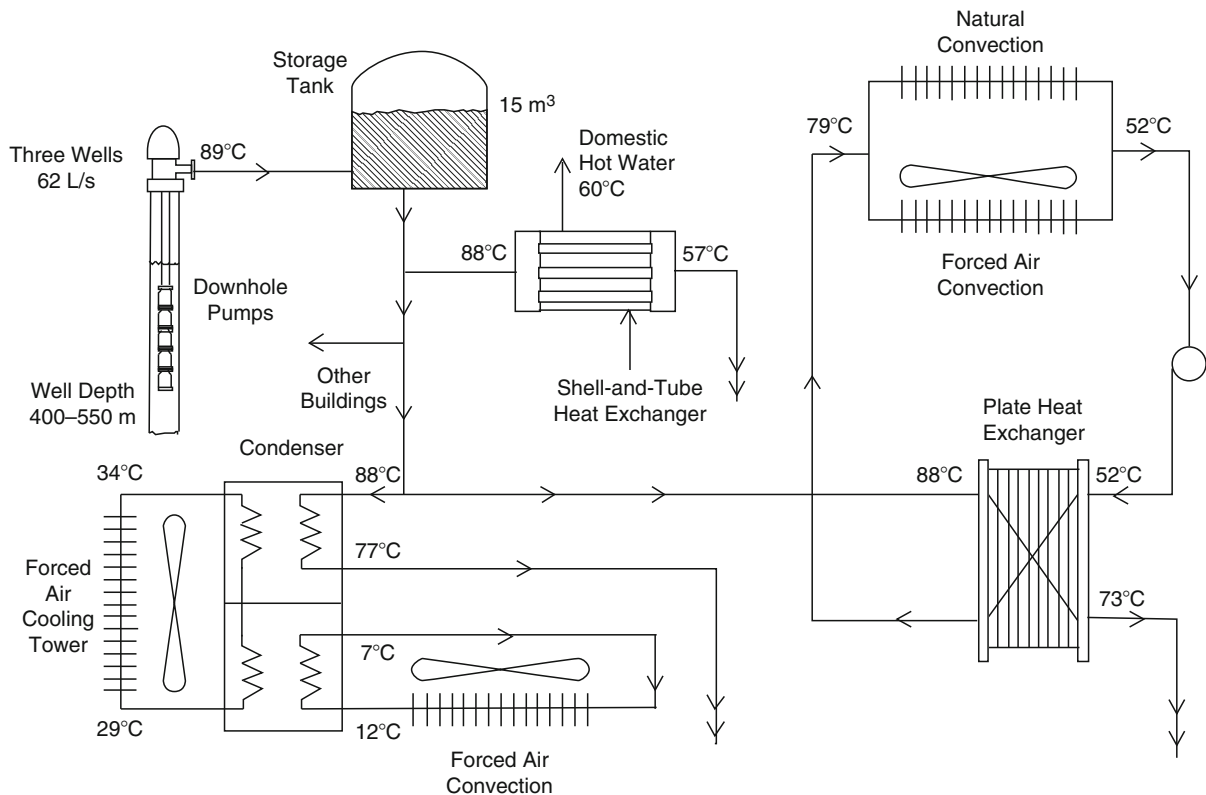
Downhole heat exchangers have been used for heating individual buildings using a closed loop of

pipe in a well extracting only heat in Klamath Falls, Oregon, Reno, Nevada, Rotorua, New Zealand, and Izmir, Turkey (see the “Heat Exchanger” section for more details). An example of space heating and cooling with low-to-moderate temperature geothermal energy is the Oregon Institute of Technology in Klamath Falls, Oregon (Fig. 5). Here, 12 buildings (approximately 70,000 m² of floor space) are heated with water from three wells at 89°C. Up to 62 L/s of fluid can be provided to the campus, with the average heat utilization rate over 0.53 MWt and the peak at 5.6 MWt. In addition, a 541 kW (154 t) chiller requiring up to 38 L/s of geothermal fluid produces 23 L/s of chilled fluid at 7°C to meet the campus cooling base load (recently decommissioned) [11, 12].

Space heating is reported in 27 countries with an installed capacity of 752 MWt and annual energy use of 9,609 TJ (2,669 GWh) based on data from country update reports presented at WGC2010 in Bali, Indonesia [4].

District Heating

District heating involves the distribution of heat (hot water or steam) from a central location through a network of pipes to individual houses or blocks of buildings. The distinction between a district heating and space heating system is that space heating usually involves one geothermal well per structure, whereas district heating involves serving multiple buildings from a well or well field. The heat is used for space heating and cooling, domestic water heating, and industrial process heat. A geothermal well field is the primary source of heat; however, depending on the temperature, the district may be a hybrid system, which would include fossil fuel and/or heat pump peaking. An important consideration in district heating projects is the thermal load density, or the heat demand divided by the ground area of the district. A high heat density, generally above 1.2 GJ/h/ha or a favorability ratio (heat load available/heat load on the system) of 2.5 GJ/ha/year, is recommended. Often fossil fuel peaking is used to meet the coldest period, rather than drilling additional wells or pumping more fluids, as geothermal can usually meet 50% of the load 80–90% of the time, thus improving the efficiency and economics of the system [13]. Geothermal district heating systems are capital intensive. The principal costs are initial investment costs for production and injection wells,

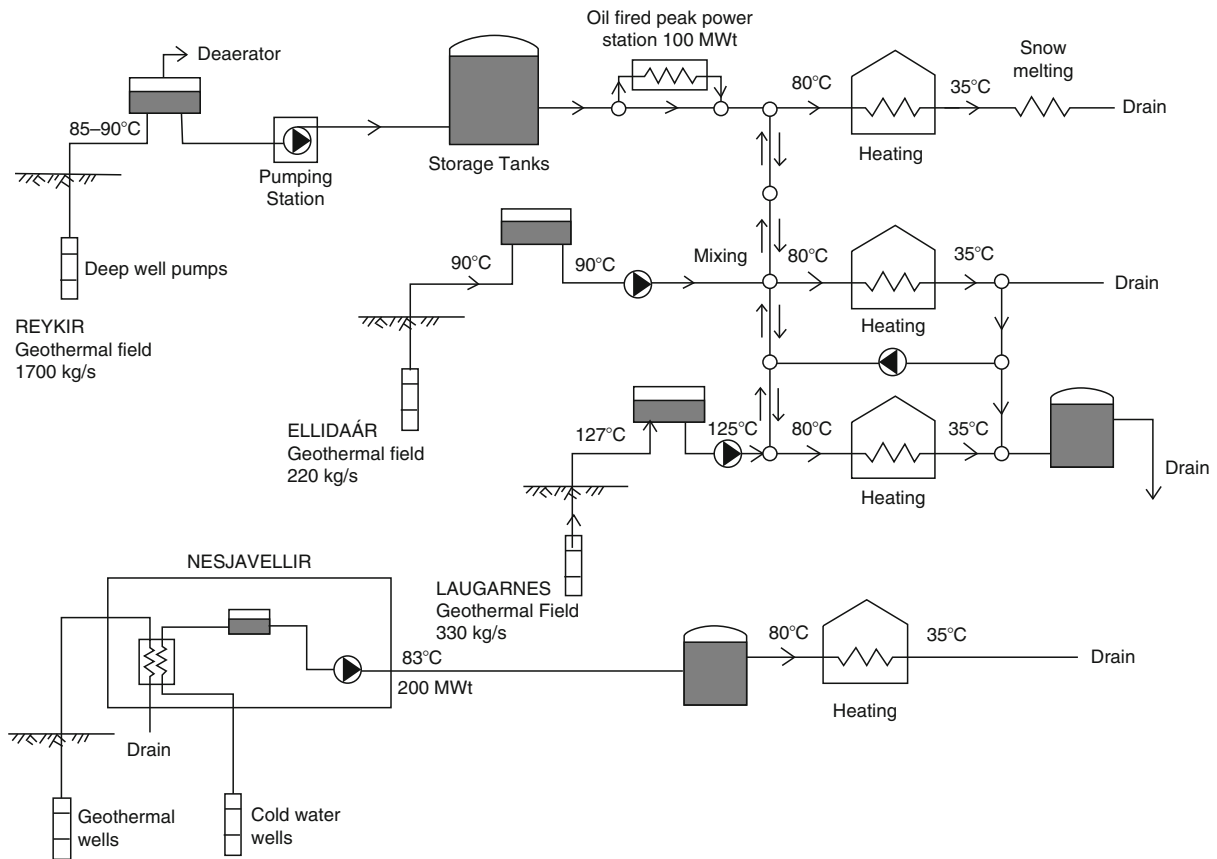


Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 5
Oregon Institute of Technology heating and cooling system

downhole and circulation pumps, heat exchangers, pipelines and distribution network, flow meters, valves and control equipment, and building retrofit. The distribution network may be the largest single capital expense at approximately 35–75% of the entire project cost. Operating expenses, however, are in comparison lower and consists of pumping power, system maintenance, control, and management. The typical savings to consumers range from approximately 30–50% per year of the cost of natural gas.

Geothermal district heating systems are in operation in 24 countries, including large installations in Iceland, France, Poland, Hungary, Turkey, Japan, China, Romania, and the USA. The Warm Springs Avenue project in Boise, Idaho, dating back to 1892 and originally heating more than 400 homes, supplies hot water or steam through a network of pipes to individual dwellings or blocks of buildings [14]. The Reykjavik, Iceland, district heating system (Fig. 6) is

probably the most famous [15, 16]. This system supplies heat for a population of around 200,000 people. The installed capacity of 1,240 MWt with peak load of 924 MWt is designed to meet the heating load to about -10°C ; however, during colder periods, the increased load is met by large storage tanks and an oil-fired booster station. The total pipeline length is 3,846 km and almost 80 million cubic meters of water are delivered annually [6]. In France, production wells in sedimentary basins provide direct heat to more than 500,000 people in 170,000 dwellings from 34 projects with an installed capacity of 300 MWt and annual energy use of 4,900 TJ/year [17]. These wells provide from 40°C to 100°C water from depths of 1,500–2,000 m. In the Paris basin, a doublet system (one production and one injection well direction drilled from on site) provides 70°C water, with the peak load met by heat pumps and conventional fossil fuel burners (Fig. 7).



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 6
Reykjavik district heating system

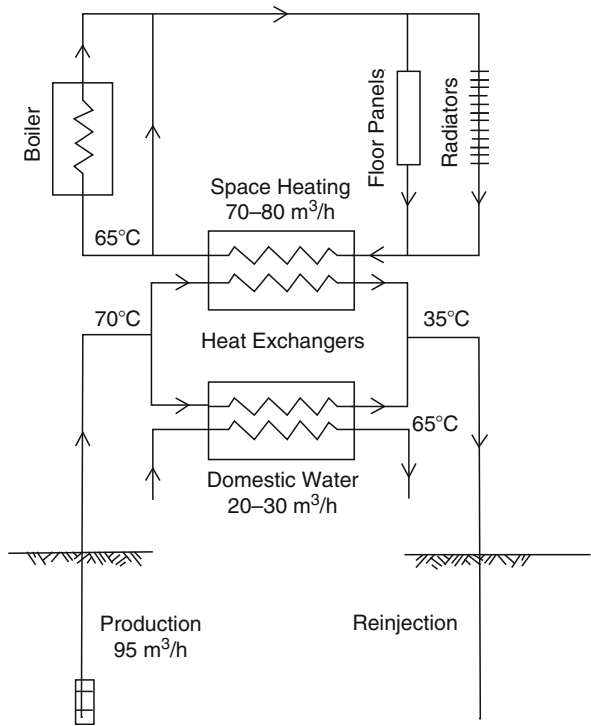
The total installed capacity for the 24 countries is 4,639 MWt and the annual energy use is 53,375 TJ (12,857 GWh) as reported in WGC2010 [4].

Agribusiness Applications

Agribusiness applications (agriculture and aquaculture) are particularly attractive because they require heating at the lower end of the temperature range where there is an abundance of geothermal resources. Use of waste heat or the cascading of geothermal energy also has excellent possibilities. A number of agribusiness applications can be considered: greenhouse heating, aquaculture and animal husbandry facilities heating, soil warming and irrigation, mushroom culture heating and cooling, and biogas generation.

Numerous commercially marketable crops have been raised in geothermally heated greenhouses in Hungary, Russia, New Zealand, Japan, Iceland, China, Tunisia, and the USA. These include vegetables, such as cucumbers and tomatoes, flowers (both potted and bedded), house plants, tree seedlings, and cacti. Using geothermal energy for heating reduces operating costs (which can account for up to 35% of the product cost) and allows operation in colder climates where commercial greenhouses would not normally be economical.

The use of geothermal energy for raising catfish, shrimp, tilapia, eels, and tropical fish has produced crops faster than by conventional solar heating. Using geothermal heat allows better control of pond temperatures, thus optimizing growth (Fig. 8). Fish breeding has been successful in Japan, China, and the

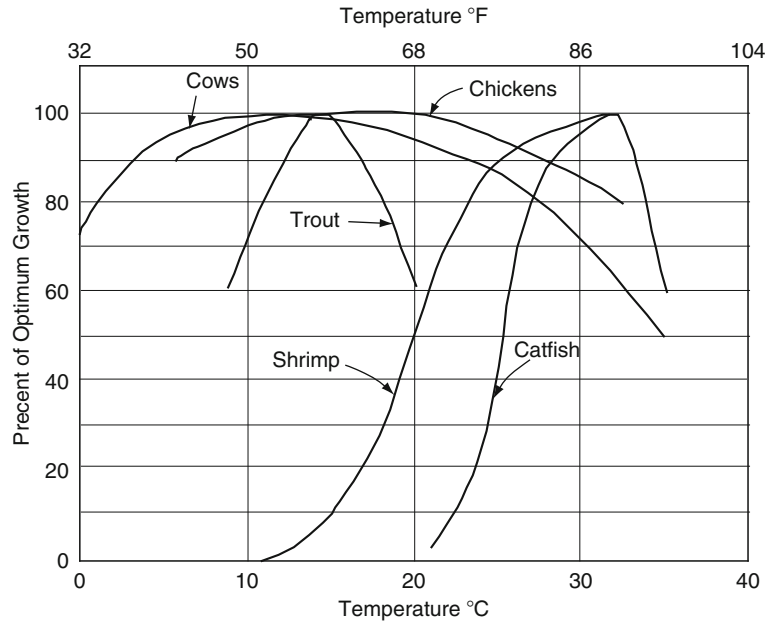


Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 7
Melun l'Almont (Paris) doublet heating system

USA. A very successful prawn raising operation, producing 400 t of Giant Malaysian Freshwater Prawns per year at US \$17 to 27/kg has been developed near the Wairakei geothermal field in New Zealand [18]. The most important factors to consider are the quality of the water and disease. If geothermal water is used directly, concentrations of dissolved heavy metals, fluorides, chlorides, arsenic, and boron must be considered, and if necessary, isolated by using a heat exchangers.

Livestock raising facilities can encourage the growth of domestic animals by a controlled heating and cooling environment. An indoor facility can lower mortality rate of newborn, enhance growth rates, control diseases, increase litter size, make waste management and collection easier, and in most cases improve the quality of the product. Geothermal fluids can also be used for cleaning, sanitizing and drying of animal shelters and waste, as well as assisting in the production of biogas from the waste.

Agribusiness uses of geothermal energy are reported in 38 countries with an installed capacity of 2,197 MWt and annual energy use of 34,785 TJ (9,662 GWh) according to WGC2010 reports [4].



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 8
Effect of temperature on animal and fish growth

Approximately two thirds of the use is for greenhouse applications, with the remaining in aquaculture production.

Industrial Applications and Agricultural Drying

Although the Lindal diagram and the current direct-use diagram (Figs. 3 and 4) shows many industrial and process applications of geothermal energy, the world's uses are relatively few. The oldest industrial use is at Larderello, Italy, where boric acid and other borate compounds have been extracted from geothermal brines since 1790. Today, the two largest industrial uses are the diatomaceous earth drying plant in northern Iceland and a pulp, paper and, wood processing plant at Kawerau, New Zealand. Notable US examples are two onion dehydration plants in northern Nevada [19], and a sewage digestion facility in San Bernardino, California. Alcohol fuel production has been attempted in the USA; however, the economics were marginal and thus this industry has not been successful. With the recent increase in fossil fuel prices, there has been renewed interest in producing ethanol and biodiesel using geothermal energy [20].

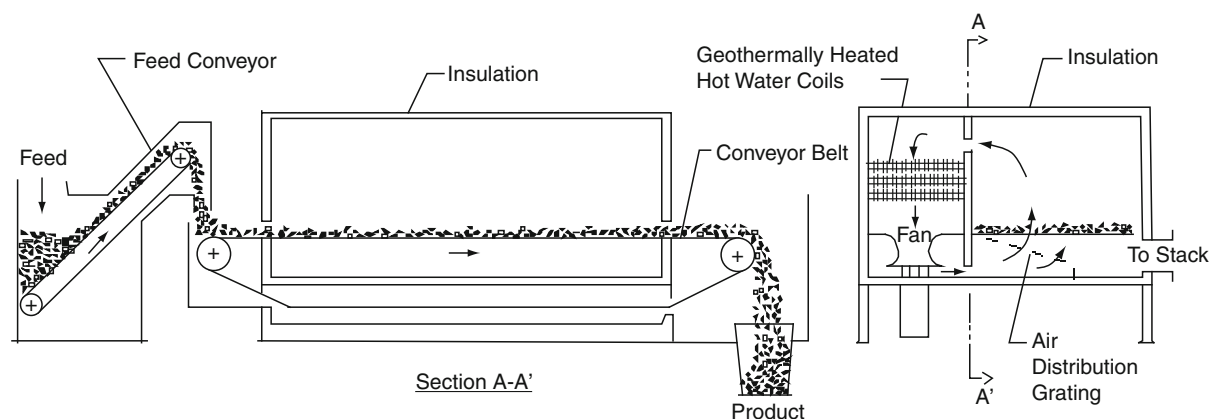
A new development in the use of geothermal fluids is the enhanced heap leaching of precious metals in Nevada by applying heat to the cyanide process [21]. Using geothermal energy increases the efficiency of the process and extends the production into the winter months.

Drying and dehydration are important moderate-temperature uses of geothermal energy. Various vegetable and fruit products are feasible with continuous belt conveyors or batch (truck) dryers with air temperatures from 40°C to 100°C as shown in Fig. 9 [22]. Geothermally drying alfalfa, onions, garlic, pears, apples, and seaweed are examples of this type of direct-use.

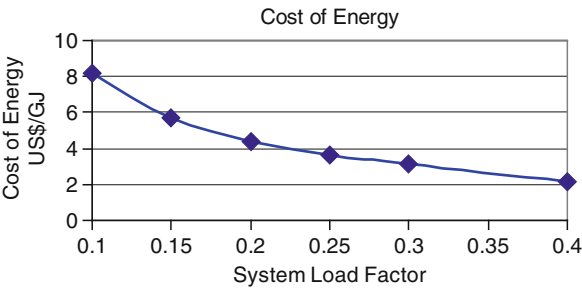
An example of a small-scale food dehydrator is one located in northeastern Greece where 4 t of tomatoes are dried annually, using 59°C geothermal water to dry 14 kg/h on racks placed in a long tunnel drier. The tomatoes are then placed in olive oil for shipment and sale. The plant is only operated by three employees. At the other end of the spectrum is the large-scale onion and garlic drying facilities located in western Nevada, USA, employing 75 workers [23]. These continuous belt drier are fed 3,000–4,300 kg/h of onions at a moisture content of around 85%, and after 24 h, produce 500–700 kg/h of dried onions at moisture contents around 4%. These large belt driers are approximately 3.8 m wide and 60 m long.

A total of 20 countries reported industrial and agricultural drying applications from WGC2010 [4], with an installed capacity of 660 MWt and annual energy use of 13,408 TJ (3,724 GWh).

Industrial applications mostly need the higher temperature as compared to space heating, greenhouses, and aquaculture projects. Examples of industrial operations that use geothermal energy are: heap leaching



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 9
Continuous belt dehydration plant, schematic



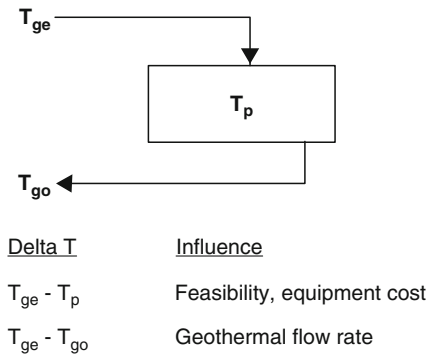
Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 10
Load factor versus cost of energy (Modified from Rafferty [24])

operations to extract precious metals in the USA (110°C), dehydration of vegetables in the USA (130°C), diatomaceous earth drying in Iceland (180°C), and pulp and paper processing in New Zealand (205°C). Drying and dehydration may be the two most important process uses of geothermal energy. A variety of vegetable and fruit products can be considered for dehydration at geothermal temperatures, such as onions, garlic, carrots, pears, apples, and dates. Industrial processes also make more efficient use of the geothermal resources as they tend to have high load factors in the range of 0.4–0.7. High load factors reduce the cost per unit of energy used as indicated in Fig. 10 [24].

Direct-Use Temperature Requirements

The design of mechanical systems involving heat transfer, such as direct-use geothermal systems, is heavily influenced by temperature. Temperature difference (delta *T* or ΔT) is particularly important as it frequently governs feasibility, equipment selection, and flow requirements for the system. Rafferty [25] addresses these issues with several “rules of thumb” that are described below. He introduces the material with the following discussion:

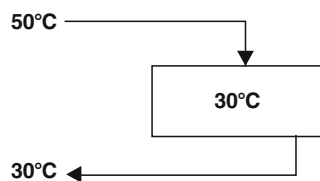
- ▶ Two primary temperature differences govern feasibility, flow requirements, and design of direct-use equipment. These are illustrated in a simplified way in Fig. 11. The first is the difference between the geothermal temperature entering the system (T_{ge}) and the process



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 11
Fundamental direct-use temperature differences [25]

temperature (T_p). This difference determines whether or not the application will be feasible. For a direct-use project, the temperature of the geothermal entering the system must be above the temperature of the process in order to transfer heat out of the geothermal water and into the process (aquaculture pond, building, greenhouse, etc.). Beyond that, it must be sufficiently above the process to allow the system to be constructed with reasonably sized heat-transfer equipment. The greater the temperature difference between the geothermal resource and the process, the lower the cost of heat exchange equipment. The key question is how much above the process temperature does the geothermal need to be for a given application.

The second temperature difference is the one between the geothermal entering the system and leaving the system (T_{ge} vs T_{go} in Fig. 11). This determines the geothermal flow rate necessary to meet the heat input requirement of the application. The greater the temperature difference between the entering and leaving temperatures, the lower the geothermal flow required. Obviously, the resource temperature is fixed. The process temperature plays a role as well since the leaving geothermal temperature cannot be lower than the process temperature to which it is providing heat. In addition, the specifics of the application and the heat transfer equipment associated with it also influence the temperature required. There are two broad groups of applications with similar characteristics in terms of heat transfer–aquaculture and pools, greenhouses, and building space heating.



Flow requirement proportional to $T_{ge} - T_{go}$
 At 40°C, flow = 2x
 At 35°C, flow = 4x
 At 32.5°C, flow = 8x

Geothermal Resources Worldwide, Direct Heat

Utilization of. Figure 12

Direct pool/pond heating (Modified from Rafferty [25])

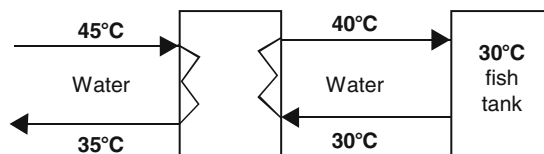
Pool and Aquaculture Pond Heating

Pond and pool heating is one of the simplest geothermal applications as it usually uses the geothermal water directly in the pond/pool to provide the required heat demand. This is illustrated in Fig. 12 [25], where 50°C geothermal water is supplied to heat the pool water to 30°C. Thus, the ΔT is 20°C, and using a flow rate of 10 L/s, the energy supplied would be 837 kW (3.0 GJ/h) ($\text{kW} = \text{L/s} \times \Delta T \times 4,184$). If the supply temperature were instead 40°C, the flow rate would have to be doubled to provide the same amount of energy, and four times at 35°C, and eight times at 32.5°C.

If the geothermal water cannot be used directly due to health restrictions, then a heat exchanger is necessary to heat treated water for the pond or pool. Following the “rule of thumb” that the heated water to the pool should be 10°C above the pool temperature, then according to the previous example, 40°C secondary water would have to be provided to the pool. Using a heat exchanger between the geothermal water and the secondary water, an additional ΔT of 5°C is required to accommodate the heat transfer between the geothermal water and the secondary water. Thus, 45°C geothermal water would be required, and on the return side of the heat exchanger, the geothermal reject fluid should be 5°C above the return temperature of the secondary water. Thus, the rule of thumb is “10/5/5” as listed below in Fig. 13.

Greenhouse and Building Space Heating

Heating of greenhouses and building often involves the transfer of heat to the air in the structure using a water-



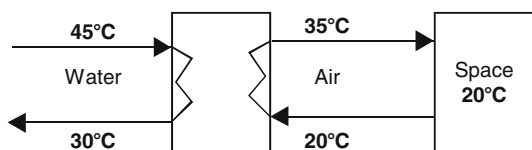
Minimum acceptable supply water temperature = process temp + 10°C
 Maximum available supply water temperature = resource temp – 5°C
 Minimum achievable geo leaving temp = process temp + 5°C

Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 13

Pond/pool heating with heat exchanger (Modified from Rafferty [25]). Minimum acceptable supply water temperature = space temperature + 15°C. Maximum available supply water temperature = geothermal water temperature – 10°C. Minimum achievable geothermal leaving temperature = return air temperature + 10°C

to-air heat exchanger, called a coil, usually consisting of finned copper tubes [25]. The simplest version of this application is shown in Fig. 14. In order to heat the space, heated air should be delivered at least 15°C above the space temperature, 20°C shown in this example. Thus, the air should be delivered at 35°C or above from the water to the coil. The reason for the large difference, 15°C, is to limit the required quantity of air circulated to meet the heating requirements at reasonable levels. Also, as the difference becomes less, the fan and duct sizes become large, and the fan power consumption can be excessive. In addition, occupant comfort is important, as when the air supply drops below the 15°C difference, the temperature of the air approaches human skin temperature, which results in a “drafty” sensation to the occupants, even at the desired air temperature. In addition, the geothermal water delivered to the water-to-air heat exchangers should be at least 10°C above the required air temperature to limit the size and cost of this heat exchanger – usually a coil type. The same ΔT is required between the leaving geothermal water and the return air temperature. Thus, to supply 20°C heat to the room, a geothermal resource temperature would have to be at least 45°C. The “rule of thumb” for this condition is then “15/10/10” as shown in Fig. 14.

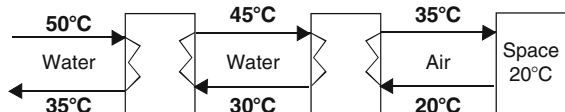
The example above assumes that the geothermal water is suitable to flow directly through the



Minimum acceptable supply water temperature = space temp. + 15°C
 Maximum available supply water temperature = geo. water temp. - 10°C
 Minimum achievable geo. leaving temperature = return air temp. + 10°C

Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 14

Space heating without isolation heat exchanger (Modified from Rafferty [25]). Supply air to space air = 15°C. Supply water to space air = 15°C. Water/air heat exchanger = supply water to supply air of 10°C. Water/water heat exchanger = supply water to supply water of 5°C



Supply air to space air = 15°C
 Water/air heat exchanger = supply water to supply air of 10°C
 Water/water heat exchanger = supply water to supply water of 5°C

Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 15

Space heating 15/10/5 rule with geothermal isolation plate heat exchanger (Modified from Rafferty [25])

water-to-air heat exchanger (coil); however, if hydrogen sulfide is present, then this gas will attack copper and solder in the coil and cause leakage and failure to the unit. Thus, in the case where the geothermal must be isolated from the heating system equipment, a plate heat exchanger is normally placed between the two circuits to protect the heating equipment [25]. A plate heat exchanger is then added to the left side of the equipment shown in Fig. 14 and resulting in the configuration shown in Fig. 15. All the temperatures shown in Fig. 14 are still valid; the difference is that the plate heat exchangers will require additional temperature input to maintain the space (home) temperature of 20°C. As in the previous example, a ΔT of 5°C is required between the geothermal supply and the output from the secondary water. Thus, the new geothermal temperature required to meet the needs of the system is 50°C. The return geothermal water can only

be cooled to 35°C as a result of the intermediate water loop return temperature of 30°C and the required 5°C ΔT . This then provides a rule of thumb of “15/10/5” as described below Fig. 15.

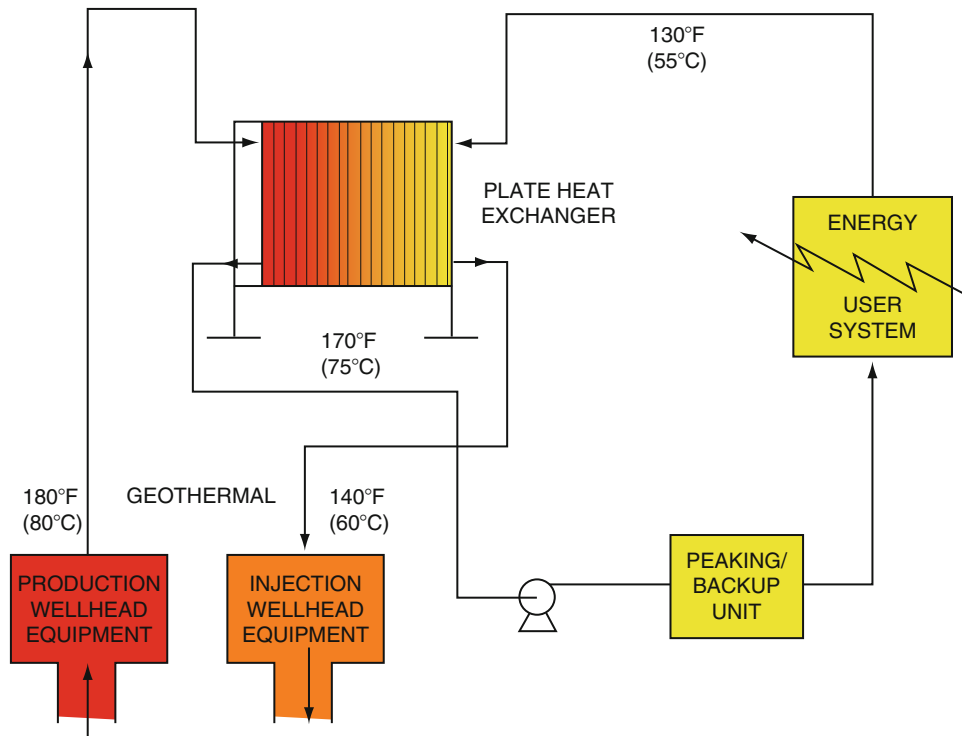
In summary, the following is provided by Rafferty [25]:

- All of the rules of thumb discussed here are exactly that. It is possible in all cases to “bend the rules,” and design systems and equipment for temperatures closer than the guidelines provided above. The values provided here are intended for initial evaluation of applications by those not in the practice of designing heating systems on a regular basis. The guidelines cited apply to new systems using commercially manufactured equipment. Homemade heat exchangers or existing equipment selected for water temperatures well above available geothermal temperature would require additional analysis.

Equipment

Standard equipment is used in most direct-use projects, provided allowances are made for the nature of geothermal water and steam. Temperature is an important consideration, so is water quality. Corrosion and scaling caused by the sometimes unique chemistry of geothermal fluids may lead to operating problems with equipment components exposed to flowing water and steam. In many instances, fluid problems can be designed out of the system. One such example concerns dissolved oxygen, which is absent in most geothermal waters, except perhaps the lowest temperature waters. Care should be taken to prevent atmospheric oxygen from entering district heating waters, for example, by proper design of storage tanks. The isolation of geothermal water by installing a heat exchanger may also solve this and similar water quality-derived problems. In this case, a clean secondary fluid is then circulated through the used side of the system as shown in Fig. 16.

The primary components of most low-temperature direct-use systems are downhole and circulation pumps, transmission and distribution pipelines, peaking or backup plants, and various forms of heat extraction equipment (Fig. 16). Fluid disposal is either surface or subsurface (injection). A peaking system may be



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 16
Geothermal direct-utilization system using a heat exchanger

necessary to meet maximum load. This can be done by increasing the water temperature or by providing tank storage (such as done in most of the Icelandic district heating systems). Both options mean that fewer wells need to be drilled thus requiring less geothermal fluid. When the geothermal water temperature is warm (below 50°C), heat pumps are often used. The equipment used in direct-use projects represents several units of operations. The major units will now be described in the same order as seen by geothermal waters produced for district heating. Detailed discussion of equipment design and use can be found in Lund et al. [26].

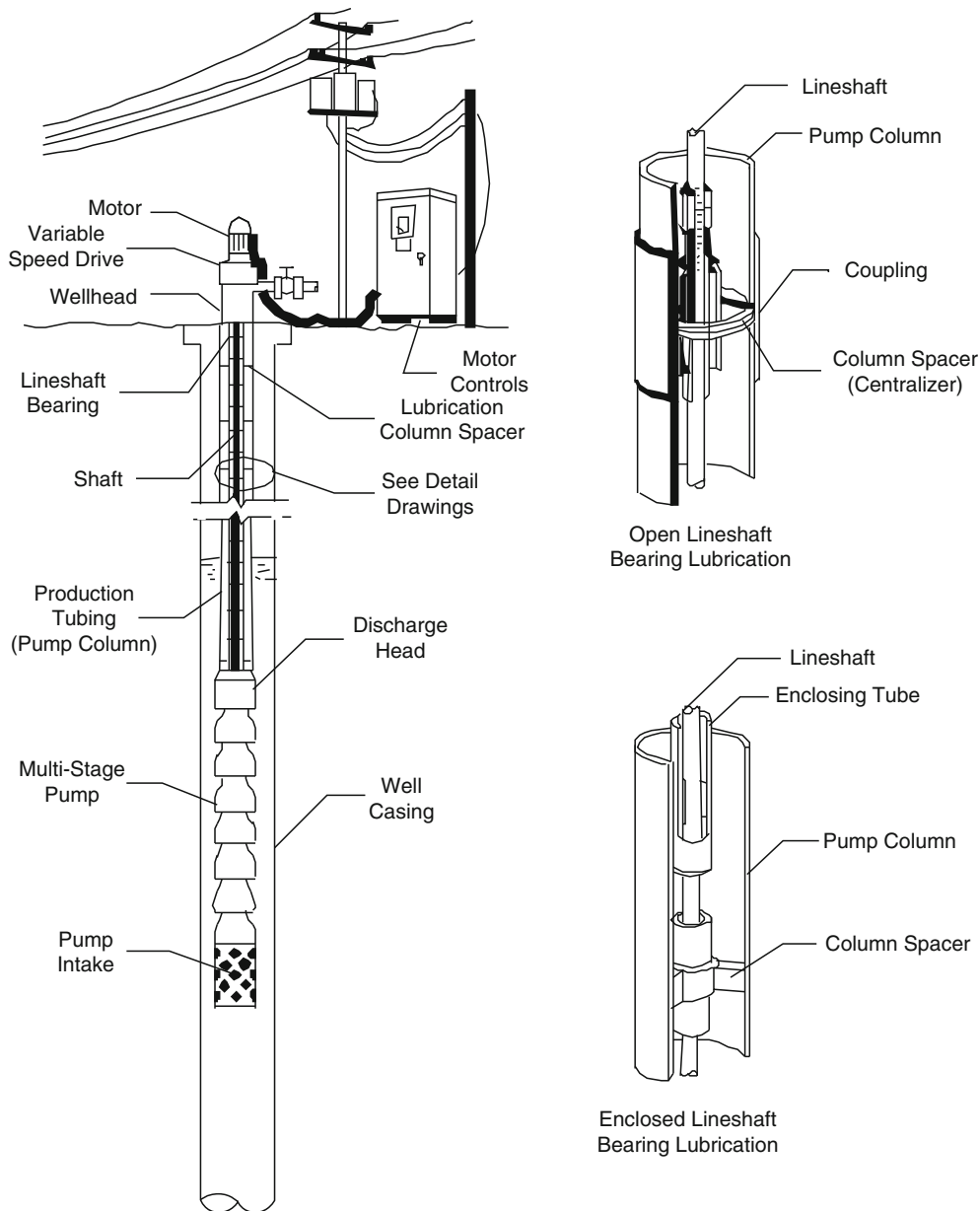
Downhole Pumps

Unless the well is artesian, downhole pumps are needed, especially in large-scale direct utilization system. Downhole pumps may be installed not only to lift fluid to the surface, but also to prevent the release of gas

and the resultant scale formation. The two most common types are: lineshaft pump systems and submersible pump systems.

The lineshaft pump system (Fig. 17) consists of a multistage downhole centrifugal pump, a surface mounted motor, and a long driveshaft assembly extending from the motor to the pump bowls. Most are enclosed, with the shaft rotating within a lubrication column which is centered in the production tubing. This assembly allows the bearings to be lubricated by oil as hot water may not provide adequate lubrication. A variable-speed drive set just below the motor on the surface can be used to regulate flow instead of just turning the pump on and off.

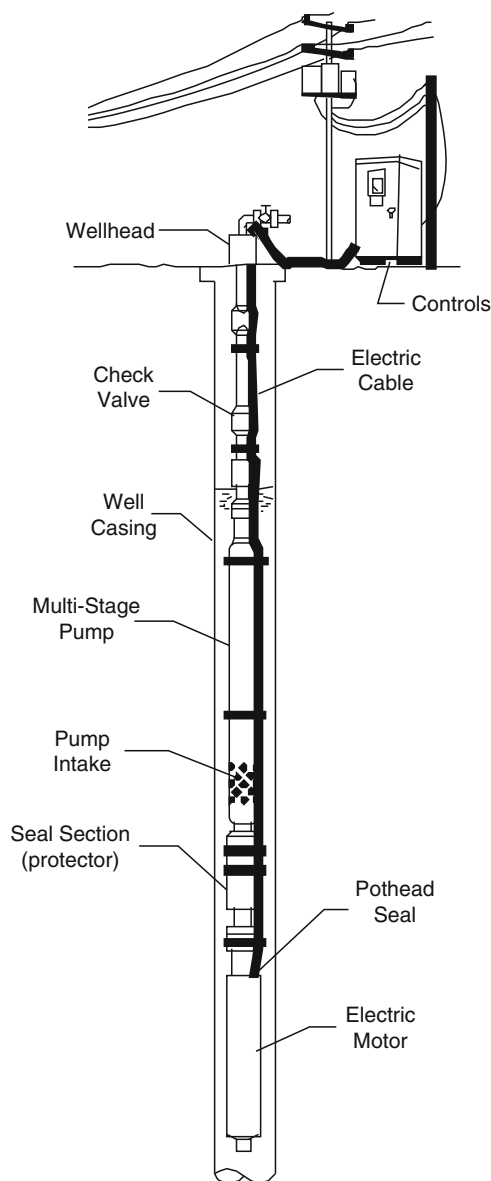
The electric submersible pump system (Fig. 18) consists of a multistage downhole centrifugal pump, a downhole motor, and a seal section (also called a protector) between the pump and motor, and electric cable extending from the motor to the surface electricity supply.



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 17
Lineshaft pump

Both types of downhole pumps have been used for many years for cold water pumping and more recently in geothermal wells (lineshafts have been used on the Oregon Institute of Technology campus in 89°C water for almost 60 years). If a lineshaft pump is used, special allowances must be made for the thermal expansion of various components and for

oil lubrication of the bearings [27]. The lineshaft pumps are preferred over the submersible pump in conventional geothermal applications for two main reasons: the lineshaft pump cost less, and it has a proven track record. However, for setting depths exceeding about 250 m, a submersible pump is required.



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 18
Submersible pump

Piping

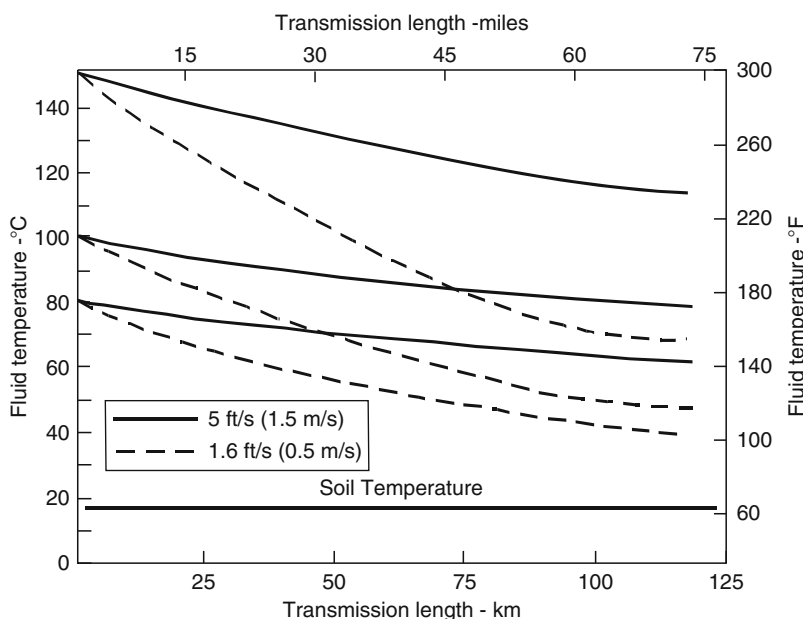
The fluid state in transmission lines of direct-use projects can be liquid water, steam vapor, or a two-phase mixture. These pipelines carry fluids from the wellhead to either a site of application or a steam-water separator. Thermal expansion of metallic pipelines heated rapidly from ambient to geothermal fluid temperatures (which could vary

from 50°C to 200°C) causes stress that must be accommodated by careful engineering design.

The cost of transmission lines and the distribution networks in direct-use projects is significant. This is especially true when the geothermal resource is located at great distance from the main load center; however, transmission distances of up to 60 km have proven economical for hot water (i.e., the Akranes project in Iceland [28], where asbestos cement covered with earth has been successful (see Fig. 20 later).

Carbon steel is now the most widely used material for geothermal transmission lines and distribution networks, especially if the fluid temperature is over 100°C. Other common types of piping material are fiberglass reinforced plastic (FRP) and asbestos cement (AC). The latter material, used widely in the past, cannot be used in many systems today due to environmental concerns; thus, it is no longer available in many locations. Polyvinyl chloride (PVC) piping is often used for the distribution network, and for uninsulated waste disposal lines where temperatures are well below 100°C. Cross-linked polyethylene pipe (PEX) have become popular in recent years as they can tolerate temperatures up to 100°C and still take pressures up to 550 kPa. However, PEX pipe is currently only available in sizes less than 5 cm in diameter. Conventional steel piping requires expansion provisions, either bellows arrangements or by loops. A typical piping installation would have fixed points and expansion points about every 100 m. In addition, the piping would have to be placed on rollers or slip plates between points. When hot water metallic pipelines are buried, they can be subjected to external corrosion from groundwater and electrolysis. They must be protected by coatings and wrappings. Concrete tunnels or trenches have been used to protect steel pipes in many geothermal district heating systems. Although expensive (generally over US \$300 per meter of length), tunnels and trenches have the advantage of easing future expansion, providing access for maintenance and a corridor for other utilities such as domestic water, waste water, electrical cables, phone lines, etc.

Supply and distribution systems can consist of either a single-pipe or a two-pipe system. The single-pipe is a once-through system where the fluid is disposed of after use. This distribution system is generally preferred when the geothermal energy is abundant and



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 19
Temperature drop in hot water transmission line

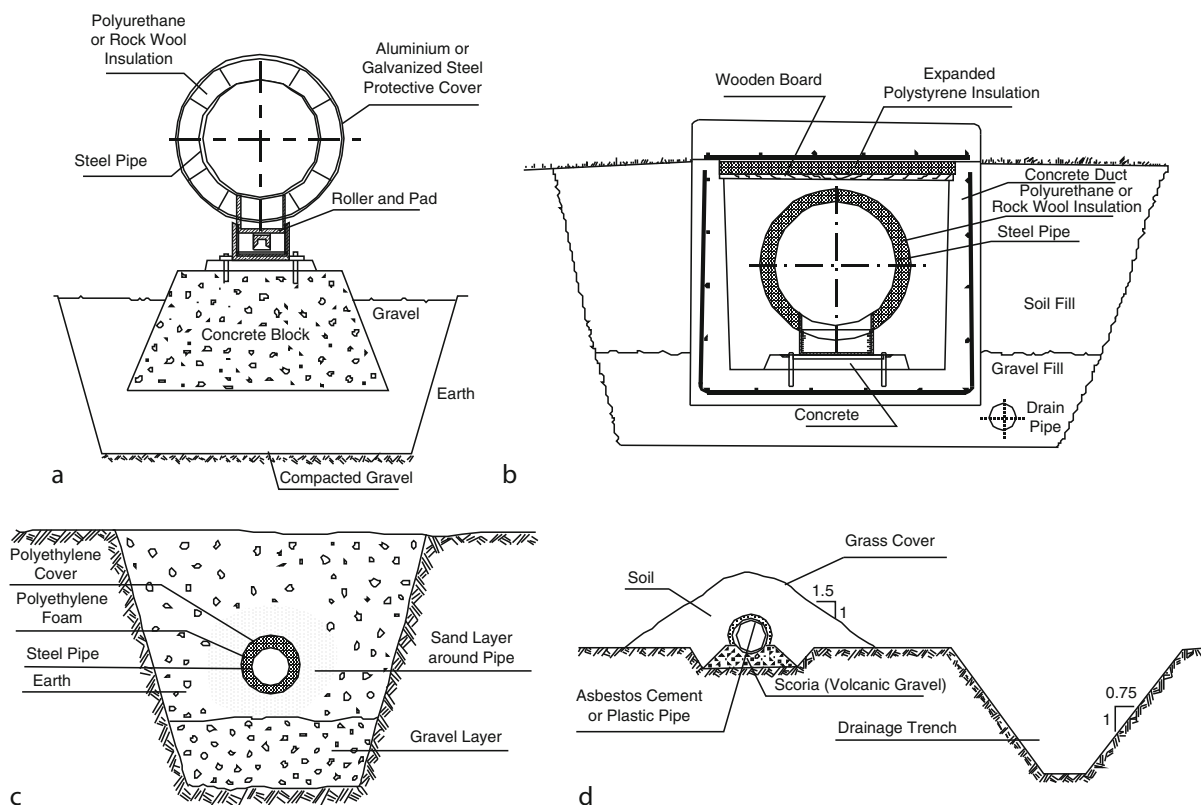
the water is pure enough to be circulated through the distribution system. In a two-pipe system, the fluid is recirculated so the fluid and residual heat are conserved. A two-pipe system must be used when mixing of spent fluids is called for, and when the spent cold fluids need to be injected into the reservoir. Two-pipe distribution systems cost typically 20–30% more than single-piped systems.

The quantity of thermal insulation of transmission lines and distribution networks will depend on many factors. In addition to minimize the heat loss of the fluid, the insulation must be waterproof and water tight. Moisture can destroy the value of any thermal insulation and cause rapid external corrosion of metallic pipe. Above ground and overhead pipeline, installations can be considered in special cases. Considerable insulation is achieved by burying hot water pipelines. For example, burying bare steel pipe results in a reduction in heat loss of about one third as compared to aboveground in still air. If the soil around the buried pipe can be kept dry, then the insulation value can be retained. Carbon steel piping can be insulated with polyurethane foam, rock wool, or fiberglass. Below-ground, such pipes should be protected with polyvinyl

chloride (PVC) jacket; aboveground, aluminium can be used. Generally, 2.5–10 cm of insulation is adequate. In two-pipe systems, the supply and return lines are usually insulated; whereas, in single-pipe systems, only the supply line is insulated.

At flowing conditions, the temperature loss in insulated pipelines is in the range of 0.1–1.0°C/km, and in uninsulated lines, the loss is 2–5°C/km (in the approximate range of 5–15 L/s flow for 15-cm diameter pipe) [29]. It is less for larger diameter pipes. For example, less than 2°C loss is experienced in the new aboveground 29 km long and 80 and 90 cm diameter line (with 10 cm of rock wool insulation) from Nesjavellir to Reykjavik in Iceland. The flow rate is around 560 L/s and takes 7 h to cover the distance. Uninsulated pipe costs about half of insulated pipe, and thus is used where temperature loss is not critical. Pipe material does not have a significant effect on heat loss; however, the flow rate does. At low flow rates (off peak), the heat loss is higher than as greater flows. Figure 18 shows fluid temperatures, as a function of distance, in a 45-cm diameter pipeline, insulated with 50 cm of urethane foam.

Several examples of aboveground and buried pipeline installations are shown in Fig. 20.



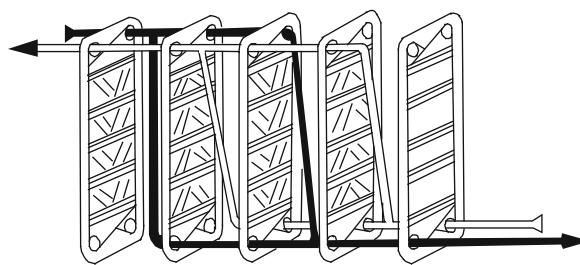
Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 20

Examples of above and below ground pipelines: (a) aboveground pipeline with sheet metal cover, (b) steel pipe in concrete tunnels, (c) steel pipe with polyurethane insulation and polyethylene cover, and (d) asbestos cement pipe with earth and grass cover

Steel piping is shown in most case, but FRP or PVC can be used in low-temperature applications. Above-ground pipelines have been used extensively in Iceland, where excavation in lava rock is expensive and difficult; however, in the USA, below ground installations are more common to protect the line from vandalism and to eliminate traffic barriers. A detailed discussion of these various installations can be found in Gudmundsson and Lund [2].

Heat Exchangers

The principal heat exchangers used in geothermal systems are the plate, shell-and-tube, and downhole types. The plate heat exchanger consists of a series of plates with gaskets held in a frame by clamping



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 21
Plate heat exchanger

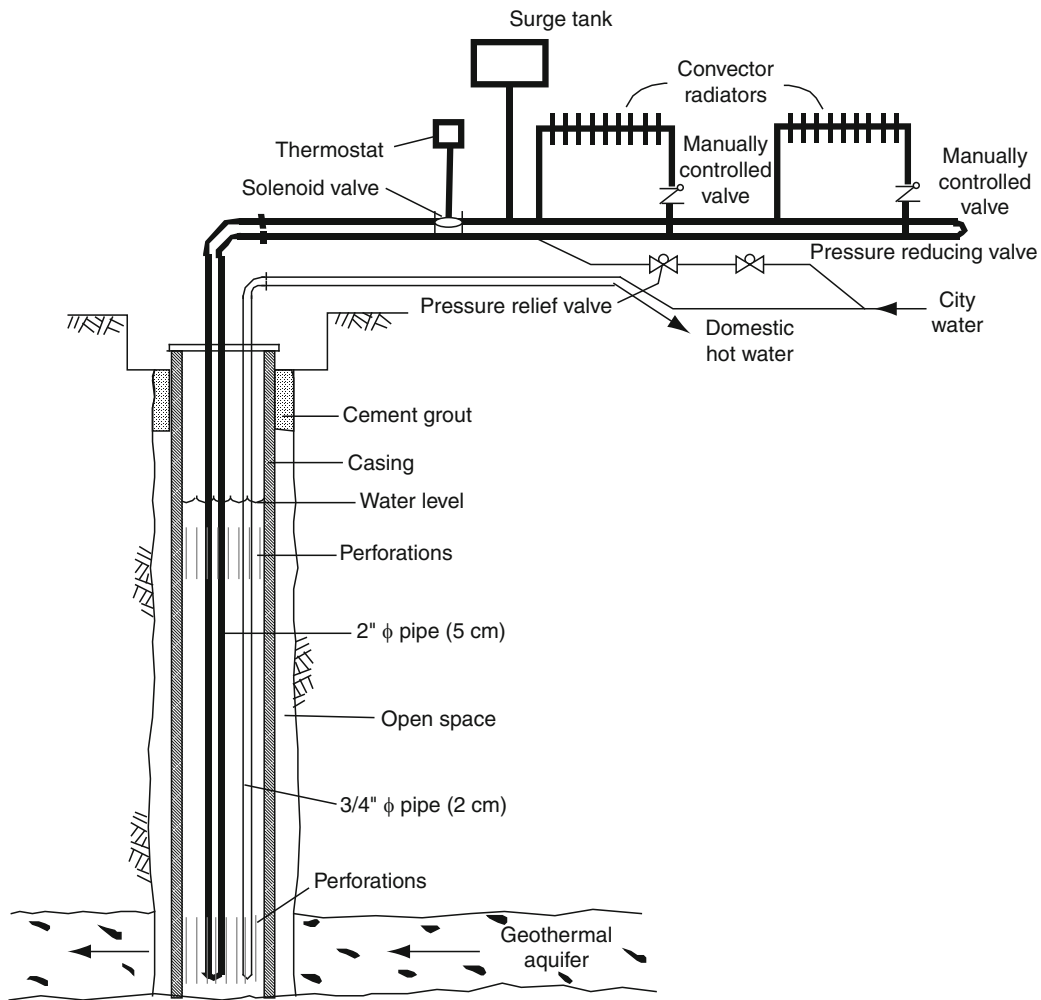
rods (Fig. 21). The countercurrent flow and high turbulence achieved in plate heat exchangers provide for efficient thermal exchange in a small volume.

In addition, they have the advantage when compared to shell-and-tube exchangers, of occupying less space, can easily be expanded when addition load is added, and cost 40% less. The plates are usually made of stainless steel; although, titanium is used when the fluids are especially corrosive. Plate heat exchangers are commonly used in geothermal heating situations worldwide.

Shell-and-tube heat exchangers may be used for geothermal applications but are less popular due to problems with fouling, greater approach temperature

(difference between incoming and outgoing fluid temperature), and the larger size.

Downhole heat exchangers eliminate the problem of disposal of geothermal fluid since only heat is taken from the well. However, their use is limited to small heating loads such as the heating of individual homes, a small apartment house or business. The exchanger consists of a system of pipes or tubes suspended in the well through which secondary water is pumped or allowed to circulate by natural convection (Fig. 22). In order to obtain maximum output, the well must be



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 22
Downhole heat exchanger (typical of Klamath Falls, Oregon)

designed to have an open annulus between the wellbore and casing, and perforations above and below the heat exchanger surface. Natural convection circulates the water down inside the casing through the lower perforations, up in the annulus, and back inside the casing through the upper perforations [30, 31]. The use of a separate pipe or promoter has proven successful in older wells in New Zealand to increase the vertical circulation [32].

Heat Pumps

At the present time, ground-coupled and groundwater (often called ground-source or geothermal) heat pump systems are being installed in great numbers in the USA, Canada, Switzerland, Sweden, Austria, and Germany [4, 33]. Groundwater aquifers and soil temperatures in the range of 5–30°C are being used in these systems. Geothermal heat pumps (GHP) utilize groundwater in wells (open loop) or by direct ground coupling (closed loop) with vertical or horizontal heat exchangers. Just about every state in the USA, especially in the midwestern and eastern states are utilizing these systems in part subsidized by public and private utilities. It is estimated that almost 3.0 million units (12 kW) are installed in 43 countries worldwide, with most in Europe, Canada and the USA. Annual growth rates are around 17%, the fastest of all the direct-use applications.

Like refrigerators, heat pumps operate on the basic principle that fluid absorbs heat when it evaporates into a gas, and likewise gives off heat when it condenses back into a liquid. A geothermal heat pump system can be used for both heating and cooling. The types of heat pumps that are adaptable to geothermal energy are the water-to-air and the water-to-water. Heat pumps are available with heating capacities of less than 3 kW to over 1,500 kW.

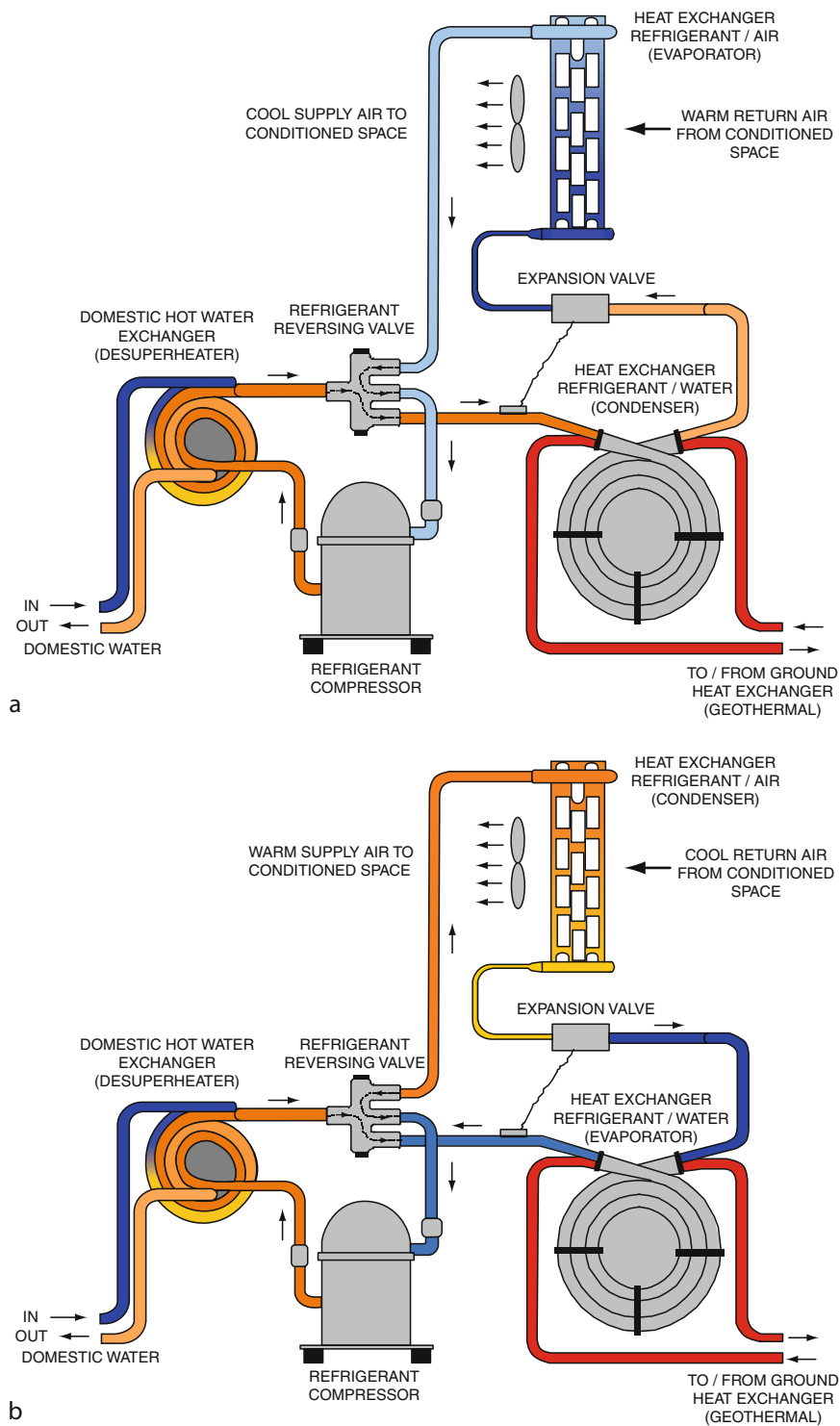
GHPs use the relatively constant temperature of the earth to provide heating, cooling and domestic hot water for homes, schools, government and commercial buildings. A small amount of electricity input is required to run a compressor; however, the energy output is in the order of four times this input. These “machines” cause heat to flow “uphill” from a lower to higher temperature location – really nothing more than

a refrigeration unit that can be reversed. “Pump” is used to describe the work done, and the temperature difference called the “lift” – the greater the lift, the greater the energy input. The technology is not new, as Lord Kelvin developed the concept in 1852, which was then modified as a GHP by Robert Webber in the 1940s. They gained commercial popularity in the 1960s and 1970s. See Fig. 23 for diagrams of typical GHP operation.

GHPs come in two basic configurations: ground-coupled (closed loop) which are installed horizontally, and vertically and groundwater (open loop) systems, which are installed in wells and lakes. The type chosen depends upon the soil and rock type at the installation, the land available and/or if a water well can be drilled economically or is already on site. As shown in Fig. 23, a desuperheater can be provided to use reject heat in the summer and some input heat in the winter for the domestic hot water heating.

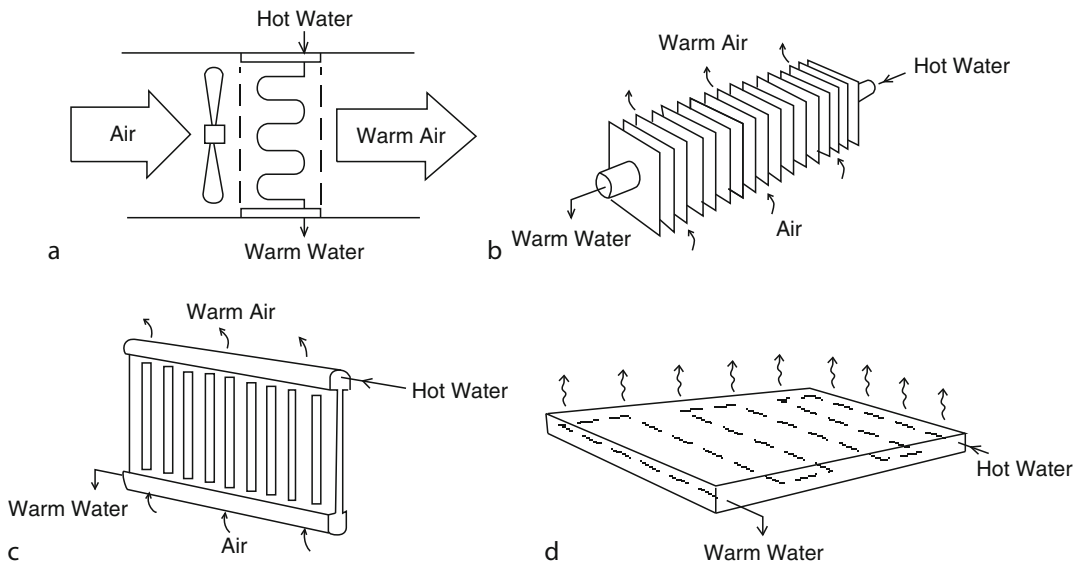
In the ground-coupled system, a closed loop of pipe, placed either horizontally (1–2 m deep) or vertically (50–100 m deep) is placed in the ground and a water-antifreeze solution is circulated through the plastic pipes (high density polyethylene) to either collect heat from the ground in the winter or reject heat to the ground in the summer [34]. The open loop system uses ground water or lake water directly in the heat exchanger and then discharges it into another well, into a stream or lake, or on the ground (say for irrigation), depending upon local laws.

The efficiency of GHP units are described by the Coefficient of Performance (COP) in the heating mode and the Energy Efficiency Ratio (EER) in the cooling mode (COP_h and COP_c , respectively, in Europe) which is the ratio of the output thermal energy divided by the input energy (electricity for the compressor) and varies from 3 to 6 with present equipment (the higher the number the better the efficiency). Thus a COP of 4 would indicate that the unit produced four units of heating energy for every unit of electrical energy input. In comparison, an air-source heat pump has a COP of around 2 and is dependent upon backup electrical energy to meet peak heating and cooling requirements. In Europe, this ratio is sometimes referred to as the



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 23

(a) GHP in the cooling cycle (From Oklahoma State University). **(b)** GHP in the heating cycle (From Oklahoma State University)



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 24

Convectors: (a) forced air, (b) material convection (finned tube), (c) natural convection (radiator), and (d) floor radiant panel

“Seasonal Performance Factor” (“Jahresarbeitszahl” in German) and is the average COP over the heating and cooling season, respectively, and takes into account system properties (see Curtis et al. [33], Lund et al. [35], and Kavanauagh and Rafferty [36] for more background material).

Convectors

Heating of individual rooms and buildings is achieved by passing geothermal water (or a heated secondary fluid) through heat convectors (or emitters) located in each room [26]. The method is similar to that used in conventional space heating systems. Three major types of heat convectors are used for space heating: (1) forced air, (2) natural air flow using hot water or finned tube radiators, and (3) radiant panels (Fig. 24). All these can be adapted directly to geothermal energy or converted by retrofitting existing systems.

Refrigeration

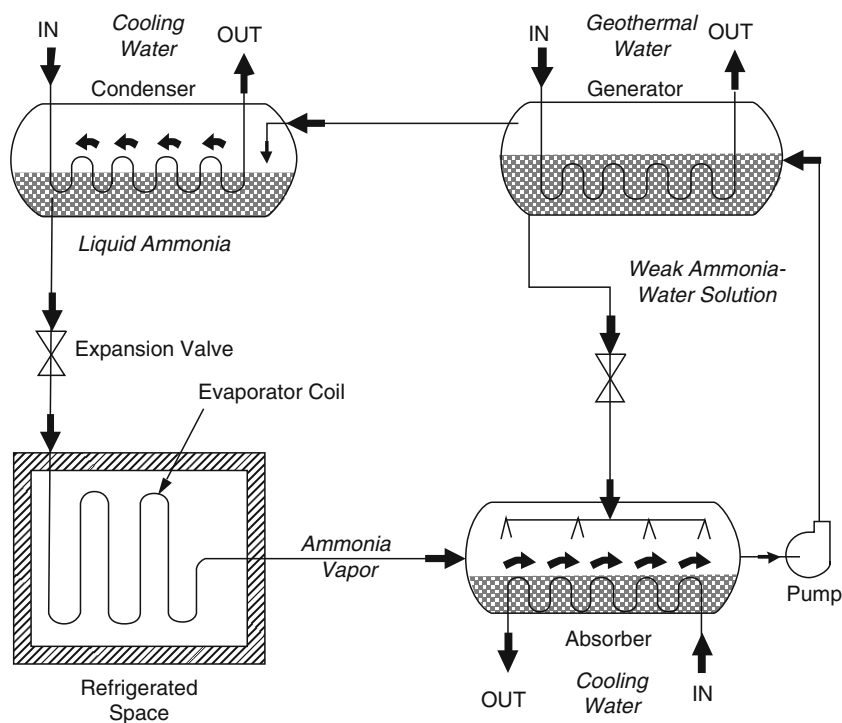
Cooling can be accomplished from geothermal energy using lithium bromide and ammonia absorption refrigeration systems [26, 37]. The lithium bromide system is the most common because it uses water as the refrigerant. However, it is limited to cooling above

the freezing point of water. The major application of lithium bromide units is for the supply of chilled water for space and process cooling. They may be either one- or two-stage units. The two-stage units require higher temperatures (about 160°C); but, they also have high efficiency. The single-stage units can be driven with hot water at temperatures as low as 77°C (such as at Oregon Institute of Technology – see Fig. 5). The lower the temperature of the geothermal water, the higher the flow rate required and the lower the efficiency. Generally, a condensing (cooling) tower is required, which will add to the cost and space requirements.

For geothermally driven refrigeration below the freezing point of water, the ammonia absorption system must be considered. However, these systems are normally applied in very large capacities and have seen limited use. For the lower temperature refrigeration, the driving temperature must be at or above about 120°C for a reasonable performance. Figure 25 illustrates how the geothermal absorption process works.

Economic Considerations

Geothermal projects require a relatively large initial capital investment, with small annual operating costs



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 25
Geothermal absorption refrigeration cycle

thereafter. Thus, a district heating project, including production wells, pipelines, heat exchangers, and injection wells, may cost several million dollars. By contrast, the initial investment in a fossil fuel system includes only the cost of a central boiler and distribution lines. The annual operation and maintenance costs for the two systems are similar, except that the fossil fuel system may continue to pay for fuel at an ever-increasing rate while the cost of the geothermal fuel is stable. The two systems, one with a high initial capital cost and the other with high annual costs, must be compared. Table 3 is an attempt to quantify the cost of various direct-use types based on experiences in the USA.

Geothermal resources fill many needs: power generation, space heating, greenhouse heating, industrial processing, and bathing to name a few. Considered individually, however, some of the uses may not promise an attractive return on investment because of the high initial capital cost. Thus, the usage of a geothermal fluid may have to be considered several times to maximize benefits. This multistage utilization, where lower and lower water temperatures are

used in successive steps, is called cascading or waste heat utilization. A simple form of cascading employs waste heat from a power plant for direct-use projects referred to as a combined heat and power application (Fig. 19) [38].

Geothermal cascading has been proposed and successfully attempted on a limited scale throughout the world. A generalized example is shown in Fig. 26. In Rotorua, New Zealand, for example, after geothermal water and steam heat a home, the owner will often use the waste heat for a backyard swimming pool and steam cooker. At the Otake geothermal power plant in Japan, about 165 t/h of hot water flows to downstream communities for space heating, greenhouses, baths, and cooking. In Sapporo, Hokkaido, Japan, the waste water from the pavement snow melting system is retained at 65°C and reused for bathing. An example of combined heat and power installation using geothermal waters down to 100°C are installed in Germany and Austria. At Neustadt Glewe in northern Germany, 98°C water from a 2,300 m-deep well at 1,700 L/s provides 11 MW (thermal) for a district

Geothermal Resources Worldwide, Direct Heat Utilization of. Table 3 Average costs of direct-use systems in the USA for 2005

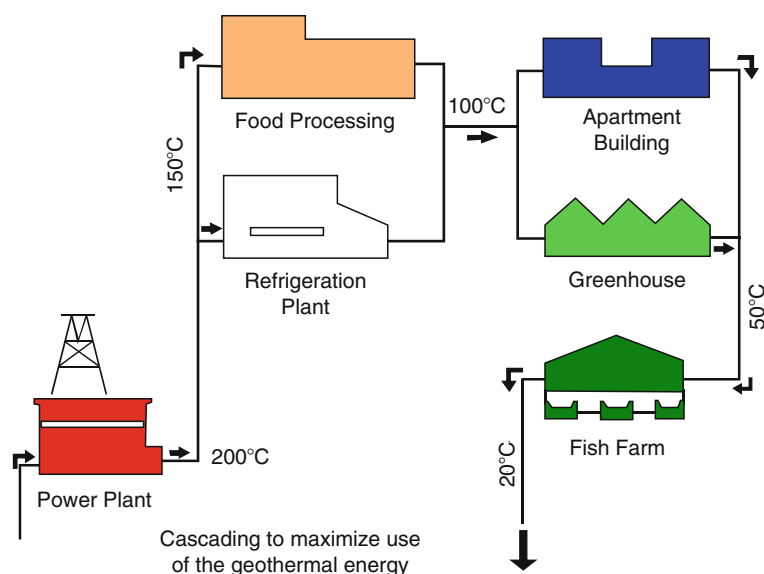
Application	Capital (\$/kW)	Cost/year (\$/kW/year)	O&M (\$/kW/year)	Total (\$/kW/year)	Capacity factor	Unit cost (cents/kWh)
Residential space heating ^a	800	71.1	7.1	78.2	0.31	3.08
Comm./inst. Space heating ^a	500	44.4	4.4	48.8	0.25	2.23
District heating	650	57.7	5.8	63.5	0.33	2.42
Greenhouse heating	250	22.2	2.2	24.4	0.26	1.11
Aquaculture pond heating	200	17.8	1.8	19.6	0.69	0.32
Geothermal heat pumps ^b	850	75.5	7.6	83.1	0.13	6.78

Based on 30-year life at 8.0% interest and O&M at 10% of capital cost

The above costs includes a shallow well (<300 m) and no retrofit costs; however, cost can vary by as much as 100% depending on the local geology, hydrology, building construction, and infrastructure

^aAssumes one production and one injection well for a single building

^bHeat pump figures are considered only for the heating mode and the capacity factor is a nationwide average



Geothermal Resources Worldwide, Direct Heat Utilization of. Figure 26

An example of cascading

heating network and 210 kW (electric) from a binary power plant meeting the electricity demands for 500 households [39].

Energy Savings

Geothermal, a domestic source of energy, could replace other forms of energy, especially fossil fuels. For many

countries, geothermal energy could lead to a reduction in their dependence on imported fuels, and for all countries, it means the elimination of pollutants such as particulates and greenhouse gases. An attempt is made here to quantify the fossil fuel savings, using a 0.35 efficiency factor if the competing energy is used to generate electricity and 0.70 if it is used directly to produce heat, such as in a furnace.

Using the 438,071 TJ/year of energy consumed in direct geothermal applications in 2010 (Table 2), and estimating that a barrel of fuel oil contains 6.06×10^9 J, and that the fuel is used to produce replacement electricity, the savings would be 206.5 million barrels of oil or 31.0 million tons of oil annually. If the oil were used directly to produce energy by burning, then these savings would be halved. The actual savings are most likely somewhere in between these two values.

The carbon savings would be 63 million tons, and the CO₂ emission savings would be 99 million tons based on using oil to produce electricity. If the savings in the cooling mode of geothermal heat pumps is considered, then this is equivalent to an additional annual savings of 101.3 million barrels (15.2 million tons) of fuel oil or 19.2 million tons of carbon pollution from burning fuel oil (see Lund et al. [4] for more details). The total of 308 million barrels (45.2 million tons) corresponds to almost 3 days of worldwide oil consumption.

There appears to be a large potential for the development of low-to-moderate enthalpy geothermal direct-use across the world which is not currently being exploited due to financial constraints and the low price of competing energy sources. Given the right environment, and as gas and oil supplies dwindle and with recent price increases, the use of geothermal energy will provide a competitive, viable, and economic alternative source of renewable energy.

Future Directions

Future development will most likely occur under the following conditions:

1. Collocated resource and uses (within 10 km apart)
2. Sites with high heat and cooling load density (>36 MWt/km²)
3. Food and grain dehydration (especially in tropical countries where spoilage is common)
4. Greenhouses in colder climates
5. Aquaculture to optimize growth – even in warm climates
6. Ground-coupled and groundwater heat pump installation (both for heating and cooling)
7. Combined heat and power installation using low-temperature resources in a binary power plant

Direct use has grown at an almost 9% annual rate over the past 10 years, and geothermal heat pumps alone has grown at a 17% annual rate over the same period [4]. The recent rise in the cost of oil and natural gas has made geothermal energy more competitive, and along with the environmental benefits associated with this renewable energy, development of this natural “heat from the earth” should accelerate in the future. At the 9% annual growth rate, the geothermal energy use should more than double over the next 10 years.

Bibliography

Primary Literature

1. Muffler LPJ (ed) (1979) Assessment of geothermal resources of the United States – 1978, USGS Circular 790, Arlington, VA, 163 p
2. Gudmundsson JS, Lund JW (1985) Direct uses of earth heat. *Int J Energy Res* 9:345–375
3. Geo-Heat Center (1997) Quarterly Bulletin 19(1), Geothermal direct-use equipment. Klamath Falls, OR, 38 p. <http://geoheat.oit.edu/bulletin/bull19-1/bull19-1.pdf>
4. Lund JW, Freeston DH, Boyd TL (2010) Direct utilization of geothermal energy 2010 worldwide review. Proceedings of the World Geothermal Congress 2010, Bali, Indonesia (CD-ROM)
5. Lund JW, Bloomquist RG, Boyd TL, Renner J (2005b) The United States of America country update – 2005. Geothermal Resources Council Transactions, vol. 29, Davis, CA (CD-ROM)
6. Ragnarsson A (2010) Geothermal development in Iceland 2005–2009. Proceedings, World Geothermal Congress 2010, Bali, Indonesia, paper no. 0124
7. Gudmundsson JS, Freeston DH, Lienau PJ (1985) The Lindal diagram. *Geothermal Resources Council Transaction* 9(1), Davis, CA, 15–19
8. Lund JW (1996) Balneological use of thermal and mineral waters in the USA. *Geothermics* 25(1), Elsevier, UK, pp 103–148.
9. Taguchi S, Itoi R, Ysa Y (1996) Beppu hot springs. *Geo Heat Cent Quart Bull* 17(2):1–6
10. Lund JW (1990) Geothermal spas in Czechoslovakia. *Geo Heat Cent Quart Bull* 12(2):20–24
11. Boyd TL (1999) The Oregon Institute of Technology Geothermal Heating System – then and now. *Geo Heat Cent Quart Bull* 20(1):10–13
12. Lund JW, Boyd T (2009) Oregon Institute of Technology Geothermal Uses and Projects, past, present and future.

- Proceedings, thirty-fourth workshop on geothermal reservoir engineering, Stanford University, Stanford, CA (CD ROM)
13. Bloomquist RG, Nimmons JT, Rafferty K (1987) District heating development guide, vol 1. Washington State Energy Office, Olympia
 14. Rafferty K (1992) A century of service: the boise warm springs water district system. *Geo Heat Cent Quart Bull* 14(2):1–5
 15. Frimannsson H (1991) Hitaveita Reykjavíkur after 60 years of operation – development and benefits. *Geo Heat Cent Quart Bull* 13(4):1–7
 16. Lund JW (2005) Hitaveita Reykjavíkur and the Nesjavellir geothermal co-generation power plant. *Geo Heat Cent Quart Bull* 26(2):19–24
 17. Boissier F, Desplan A, Laplaige P (2010) France country update. Proceeding of the World Geothermal Congress 2010, Bali, Indonesia, paper no.161
 18. Lund JW, Klein R (1995) Prawn park – Taupo, New Zealand. *Geo Heat Cent Quart Bull* 16(4):27–29
 19. Lund JW (1995) Onion dehydration. *Geothermal Resources Council Transaction*, vol. 19, Davis, CA, 69–74
 20. Chiasson A (2007) Geothermal energy utilization in ethanol production. *Geo Heat Cent Quart Bull* 28(1):2–5
 21. Trexler DT, Flynn T, Hendrix JW (1990) Heap Leaching. *Geo Heat Cent Quart Bull* 12(4):1–4
 22. Lund JW, Rangel MA (1995) Pilot fruit drier for the Los Azufres geothermal field, Mexico. *Proceedings of the World Geothermal Congress 1995*, 2335–2338
 23. Lund JW, Lienau PJ (1994) Onion dehydration. *Geo Heat Cent Quart Bull* 15(4):15–18
 24. Rafferty K (2003) Industrial process and the potential for geothermal applications. *Geo-Heat Center Quart Bull* 24(3):7–12
 25. Rafferty K (2004) Direct-use temperature requirements: a few rules of thumb. *Geo-Heat Center Quart Bull* 25(2):1–3
 26. Lund JW, Lienau PJ, Lunis BC (eds) (1998) Geothermal direct-use engineering and design guidebook. *Geo-Heat Center, Klamath Falls*, p 470
 27. Rafferty K, Keiffer S (2002) Thermal expansion in enclosed lineshaft pump columns. *Geo Heat Cent Quart Bull* 23(2):11–15
 28. Ragnarsson A, Hrólfsson I (1998) Akranes and Borgarfjörður district heating system. *Geo Heat Cent Quart Bull* 19(4):10–13
 29. Ryan GP (1981) Equipment used in direct heat projects. *Geothermal Resources Council Transactions*, vol. 5, Davis, CA, pp 483–485
 30. Culver GG, Reistad GM (1978) Evaluation and design of downhole heat exchangers for direct applications. *Geo-Heat Center, Klamath Falls*
 31. Geo-Heat Center (1999) Downhole heat exchangers. *Geo-Heat Center Quart Bull* 20(3):28 p. <http://geoheat.oit.edu/bulletin/bull20-3/bull20-3.pdf>
 32. Dunstall MG, Freeston DM (1990) U-tube downhole heat exchanger performance in a 4-in. well, Rotorua, New Zealand. *Proceedings of the 12th New Zealand Geothermal Workshop, Auckland, New Zealand*, pp 229–232
 33. Curtis R, Lund J, Sanner B, Rybach L, Hellström G (2005) Ground source heat pumps – geothermal energy for anyone, anywhere: current worldwide activity. *Proceedings of the World Geothermal Congress, 2005 (CD-ROM)*, International Geothermal Association, Antalya
 34. Rafferty K (2008) An Information survival kit for the prospective geothermal heat pump owner. *HeatSpring Energy, Cambridge, MA*, p 32
 35. Lund JW, Sanner B, Rybach L, Curtis R, Hellström G (2003) Ground-source heat pumps – a world overview, *Renewable Energy World*. James & James, London, pp 218–227
 36. Kavanaugh S, Rafferty K (1997) Ground-source design of geothermal systems for commercial and institutional buildings. *ASHRAE, Atlanta*, p 167
 37. Rafferty K (1983) Absorption refrigeration: cooling with hot water. *Geo Heat Cent Quart Bull* 8(1):17–20
 38. Geo-Heat Center (2005) Combined heat and power plant. *Geo-Heat Center Quart Bull* 26(3):36. <http://geoheat.oit.edu/bulletin/bull26-3/bull26-3.pdf>
 39. Lund JW (compiled by) (2005b) Combined heat and power plant, Neustadt-Glewe, Germany. *Geo-Heat Center Quart Bull* 26(2):31–34

Books and Reviews

- Cataldi R, Hodgson SF, Lund JW (eds) (1999) *Stories from a heated earth – our geothermal heritage*. International Geothermal Association and the Geothermal Resources Council, Davis, p 569
- Kavanaugh SP, Rafferty K (1997) *Ground-source heat pumps – design of geothermal systems for commercial and institutional buildings*. American Society of Heating Refrigerating and Air-Conditioning Engineers, Atlanta, p 167
- Lund JW (1996) *Lectures on direct utilization of geothermal energy*, United Nations University, Geothermal Training Program, Report 1, Orkustofnun, Reykjavik, Iceland, 123 p
- Lund JW, Lienau PJ, Lunis BC (eds) (1998) *Geothermal direct-use engineering and design guidebook*. Geo-Heat Center, Oregon Institute of Technology, Klamath Falls, p 454

Websites

- European Geothermal Energy Council, Belgium, www.geothermie.de/egec_geothernet/menu/frameset.htm
- Geo-Heat Center, Oregon Institute of Technology, <http://geoheat.oit.edu>
- Geothermal Education Office, USA, <http://geothermal.marin.org>
- IEA (International Energy Agency) Heat Pump Center, The Netherlands, www.heatpumpcentre.org
- International Ground Source Heat Pump Association, USA, <http://www.igshpa.okstate.edu>

Geothermal Resources, Drilling for

JOHN T. FINGER

Sandia National Laboratories, retired, Albuquerque, NM, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Well Cost

Planning and Designing the Well

Drilling System Selection Criteria

Drill Bits and Bottom-Hole Assembly

Drilling Fluids

Lost Circulation

Well Control

Completions and Cementing

Instrumentation (Drilling and Mud Logging)

Future Directions

Bibliography

Glossary

Barrel An extremely common unit of volume in the drilling industry, equal to 42 US gallons or 178 l.

BHA (bottom-hole assembly) The assembly of heavy drilling tools at the bottom of the drill string; normally includes bit, reamers, stabilizers, drill collars, heavy-weight drill pipe, jars, and other miscellaneous tools.

Blow out Uncontrolled flow of fluids from a wellhead or wellbore.

BOP (blow-out preventer) One or more devices used to seal the well at the wellhead, preventing uncontrolled escape of gases, liquids, or steam; usually includes annular preventer (an inflatable bladder that seals around drill string or irregularly shaped tools) and rams (pipe rams or blind rams: pipe rams seal around the drill pipe if it is in the hole, blind rams seal against each other if the pipe is not in the hole).

Dewar A double-walled container or heat shield, similar to a vacuum flask, which insulates a piece of equipment from high temperature.

Directional drilling Deliberately drilling on a controlled non-vertical trajectory, usually done to improve productivity.

Drill collars Heavy-walled sections at the bottom of the drill string; provide stiffness, vibration control, and most of the weight on the bit.

Fish Any part of the drill string, or other tools, accidentally left in the hole; also, *fishing* – trying to retrieve a fish.

H₂S (hydrogen sulfide) A poisonous gas sometimes found in geothermal drilling.

LCM (lost-circulation material) Any material used to plug formation fractures to avoid loss of drilling fluid.

Stand More than one joint of drill pipe screwed together; when tripping, pipe is handled in stands to avoid making and breaking every connection – for a coring rig, a typical stand is four 3 m joints (12 m), but for a large rotary rig, a stand is three 10 m joints (30 m).

Sub Generic name for part of the drill string; for example, instrumentation sub carries instruments for navigation or logging; crossover sub allows different threads to be connected; bent sub forms a slight angle between the axis of the drill string and the axis of a downhole motor, allowing directional drilling.

Trip Any event of pulling the drill string out of the hole and returning it.

Twist-off Failure mode in which some element of the drill string parts, leaving at least one portion of the drill string in the hole.

Under-pressured Describes the pore pressure of in situ fluids during drilling as less than the static head of a water column to the same depth in the wellbore.

Washout A hole or leak in the drill string; often caused by fatigue failure, but very dangerous because the flow of high-pressure drilling fluid through the leak will quickly enlarge it to the point of parting the drill string.

Definition of the Subject

The word “geothermal” comes from the combination of the Greek words *gê*, meaning Earth, and *thérn*, meaning heat. Quite literally, geothermal energy is the heat of the Earth. Geothermal resources are concentrations of the Earth’s heat that can be extracted economically for some useful purposes. All existing applications of geothermal energy use a circulating fluid to carry the heat from depth

to its use at the surface, and this means that holes must be drilled for access to or introduction of these fluids. Drilling, therefore, is a major component of any geothermal project's development.

This entry describes the overall process of drilling, with emphasis on the ways in which geothermal drilling differs from other kinds of drilling, such as that for oil and gas. The entry also focuses on the drilling of relatively large-diameter, high-temperature holes, such as those most often used to supply electrical generating plants, and specifically does not address the following topics: low-temperature drilling for direct-use application, well maintenance and workover, or drilling for geothermal heat-pump installations. The entry should by no means be considered a set of instructions on how to drill a geothermal well, but is intended to illuminate some of the major decisions that will be necessary during that process.

Introduction

Geothermal energy is a growing enterprise. Worldwide electricity production increased from 6,833 MWe (megawatts electric) in 1995 to 9,966 MWe in 2008, and direct use in 2005 displaced more than thirty million barrels of oil [1]. In spite of this growth, geothermal drilling activity is minuscule compared to oil and gas – fewer than 100 geothermal wells were drilled in the USA during 2008, while the total for oil and gas exceeded 50,000 [2]. This

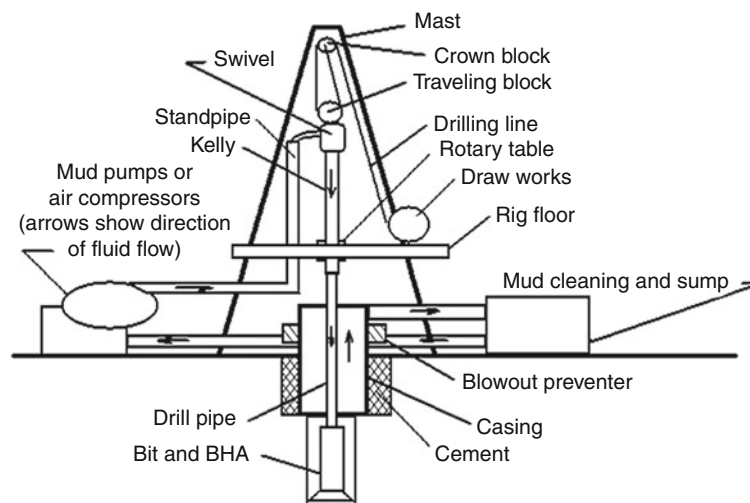
means that few service companies or drilling contractors can sustain their business solely within the geothermal industry, and it also leads to a lack of research into tools or techniques specifically aimed at geothermal drilling. A substantial number of deep gas wells, however, now have producing horizons with geothermal-like temperatures ($\sim 175^{\circ}\text{C}$), so this has brought new interest into high-temperature drilling.

Before describing the aspects that make geothermal drilling unique, however, a brief summary of the fundamental process will be useful. The process of drilling, rather than digging, holes in the ground has been under development for thousands of years, but the techniques we now know as “conventional rotary drilling” began to be developed around the end of the nineteenth century. This technology, with only minor variations, is ubiquitous in the oil, gas, geothermal, water well, and mining industries. There is an extensive literature on the principles and practices of this kind of drilling [3], and a baseline system – a tall, steel derrick supporting a string of pipe which turns a bit to drill the hole – is at least superficially familiar to most readers (Fig. 1).

Basic Drilling Functions

Any drilling system must perform six basic functions:

1. Transmit energy from the surface to the rock face
2. Reduce rock from its more-or-less monolithic state



Geothermal Resources, Drilling for. Figure 1
Drill rig diagram

3. Remove the reduced rock from the wellbore
4. Maintain control of any pressures encountered in the wellbore
5. Keep the hole open, stable at some minimum diameter, and on the desired trajectory while drilling
6. Preserve and control the well for some indefinite, but relatively long, time

Each of these functions is described in more detail, with the same numbering reference as above. In the baseline system, all of the equipment necessary for the drilling operation is organized around the derrick, or mast. This is a steel tower, ranging from 16 m to 50 m in height, which supports the drill pipe with the bit and all the other downhole equipment, and which provides a platform for much of the other equipment necessary to drill the hole. Every rig, except for the smallest ones, has a floor just above ground level where most activity required to operate the rig takes place. The driller, who has minute-by-minute control of the rig's operation, has a control console here and most equipment handling (adding a new piece of drill pipe, making and breaking drill string connections, changing bits, etc.) takes place on the floor. In smaller rigs, the mast and the floor are a unit and are simply raised into position in preparation for drilling. Because of larger hole sizes, geothermal wells usually need bigger rigs, which may require 50–60 large truck loads for transportation. These rigs are usually assembled at the drill site, a job which may take several days, even in accessible locations on land.

1. To make the hole, energy must be transmitted from the surface to the rock face at the end of the wellbore. Power supply for drilling has evolved from the early days of steam-driven, mechanically coupled rigs to the current standard of diesel-electric drive. In this configuration, two to four diesel engines (up to 1,500 kW each) drive electric generators, which supply power to individual electric motors driving the rotary table, drawworks, mud pumps, and other equipment. The rotary table is a mechanism, usually inset into the rig floor, which turns the drill string to break rock and advance the hole. (A “drill string” comprises the drill pipe plus the bottom-hole assembly, or BHA. The BHA includes drill collars, stabilizers, bit, and any other specialized tools below the drill pipe.) Torque is applied to the kelly, which is attached to the top of the drill string. The kelly

is a section of pipe with a square or hexagonal outside cross section that engages a matching bushing in the rotary table. This bushing lets the rotary table continuously turn the kelly and drill string while they slide downward as the hole advances.

The upper end of the kelly is attached to a “swivel,” which is a rotating pressure fitting that allows the drilling fluid to flow from the mud pumps, up the standpipe, through the kelly hose, into the swivel, and finally down the drill pipe as it rotates. The swivel is carried by the hook on the traveling block and it suspends most of the weight of the drill string while drilling.

Moving the drill string into and out of the hole is called tripping. Trips are usually required when the bit or some other piece of downhole equipment must be replaced, or because of some activity such as logging, testing, or running casing. Clearly, trips take longer as the hole grows deeper. Raising or lowering the drill string for a trip is done by the drawworks, which is a large winch. The drawworks reels in or pays out a wire rope (drilling line) that passes over the crown block at the top of the rig's mast and then down to the traveling block which carries the hook, which in turn suspends the drill string or casing. Depending on what mechanical advantage is required, the drilling line is reeved several times between the crown and traveling blocks, as in a block and tackle.

2. Attached to the bottom of the drill string, the bit rotates to break (reduce) the rock from its more-or-less monolithic state into small fragments (usually called “cuttings”) and to advance the hole. A tremendous variety of bits is available, and some of the important types are discussed in more detail in the section [Drill Bits and Bottom-Hole Assemblies](#).

3. Once the rock has been reduced to chips and fines, it must be removed from the hole bottom to expose fresh rock surface and to avoid wasting energy by re-grinding these same cuttings. This cleaning is done by a stream of fluid that circulates down the drill pipe, passes through ports (called “jets”) in the bit, and returns up the annulus between the wellbore wall and the outside of the drill string, carrying the rock cuttings back to the surface. This fluid is sometimes a gas (air, nitrogen, and natural gas), but is most often a liquid, universally known as “mud” from its origin as a mixture of water and clay.

Air drilling, in which the hole is cleaned by a compressor-driven airstream, generally makes hole faster than mud drilling, but suffers severe issues with well control, hole stability, drill-pipe erosion, and difficulty with handling water influx. Mud drilling uses pumps to circulate the liquid, which not only carries cuttings but stabilizes the wellbore and lubricates the bit and drill string. When mud returns to the surface, it is cleaned to remove most of the rock cuttings and is then recirculated. Pumping mud while drilling, at typical flow rates of 12 to 50 l/s, with pressures up to 20 MPa, can represent more than 75% of the rig's total power consumption. Requirements for drilling fluids and the circulating system are described in more detail in the Section [Drilling Fluids](#).

4. During drilling, the personnel and equipment must be protected against unexpected pressure surges in the wellbore. In oil and gas drilling, these surges can come from hydrocarbon fluids trapped under impermeable rock which holds them at pressures higher than the static head of the fluid column in the wellbore, and in geothermal operations the surges come from hot formations which heat the pore or wellbore fluids above the saturation temperature at the static wellbore pressure. In either case, the first line of control is the weight of the fluid column in the wellbore. With a gas column, this weight is negligible, but with mud the liquid density will range from slightly greater than water (specific gravity ~ 1.05) to almost three times that. In addition to the clays and additives that raise the viscosity of the mud to improve hole cleaning, weighting materials such as barite are often added to increase the mud's density and enable it to control higher downhole pressures. If a pressure surge cannot immediately be controlled with fluid weight, the wellbore can be mechanically sealed at the surface with BOPs, or blow-out preventers. See the Section [Well Control](#) for more detailed information on blow-out prevention equipment and well control.

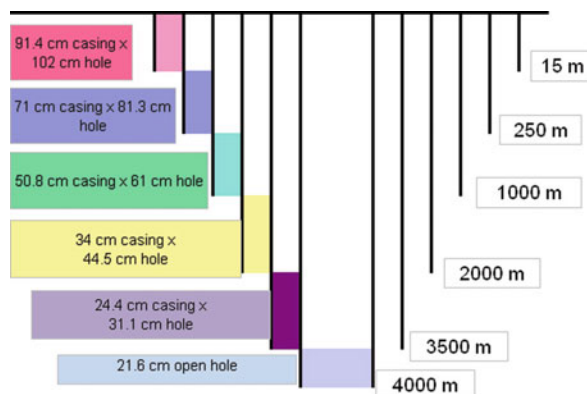
5. Hole stability can be a problem after drilling through some kinds of rock for several reasons: the formation may tend to swell (because it absorbs water) or squeeze (extrude into the wellbore because of overburden pressure), both of which reduce the hole diameter; or chunks of the wellbore wall may cave or slough into the hole. These phenomena can cause problems ranging from minor (the necessity to clean

out debris or to ream part of the hole) to major (stuck drill string). With gas drilling, there is no liquid to cause swelling, but there is no fluid pressure to counteract squeezing. With mud, these problems can often be eliminated or mitigated by the pressure of the fluid column or by the mud's chemical composition.

Controlling the hole's trajectory has two major components: keeping it either straight or, in the case of directional drilling, deviating along a relatively smooth curve; and making sure that the hole is advancing in the proper direction. Keeping the hole straight or smoothly curved is necessary to allow the casing to be run easily; getting a string of casing stuck during its deployment can be a serious and expensive problem. If the hole is being directionally drilled, the two principal aspects of hole trajectory are inclination and azimuth. Vertical holes depend mostly on the pendulum effect of gravity to keep the drill string pointed downward, but sometimes the combination of BHA design and formation properties will drive the hole away from verticality.

If the hole trajectory must be changed, either to correct unwanted deviation or to steer it toward a specific target, it is necessary to force the hole into the correct trajectory with directional drilling. Directional drilling is extensively used in the oil and gas industry to increase productivity by keeping the hole in hydrocarbon-bearing strata, and similarly in geothermal reservoirs to intersect more productive fractures. Directional drilling is a complex topic [4] and there are a number of techniques available for performing it, but a complete discussion is far beyond the scope of this entry. One aspect that is relevant to geothermal drilling is that the drilling motors and electronic steering-survey tools are susceptible to high temperature, so the hole trajectory is usually set in the upper, cooler interval of the hole and then efforts are just to keep it straight from there.

6. Once the hole is drilled to the target depth, it must be kept open for testing or production. This is conventionally done by putting steel pipe, or casing, into the hole and cementing it in place. Casing is not done all at once, at the end of drilling, but is placed sequentially in the hole as it reaches increasingly greater depths. As each casing string is placed and cemented, the hole interval below that string must be smaller than the one above, since the new drill bit must pass through the casing just set. The completed hole, then,



Geothermal Resources, Drilling for. Figure 2
Diagram of typical casing program

will usually have two to four concentric strings of casing cemented in place with an open-hole section at the bottom for production of the desired fluids (Fig. 2).

To complete any given interval of the well, casing (which is several centimeters smaller than the hole diameter at that point) is lowered almost to the bottom of the hole; then cement is pumped down the inside of the casing and displaced with mud up the annulus between the casing and the wellbore wall. Because large volumes of cement must be pumped quickly, and at high pressure because of the density difference between the mud and cement, specialized cementing equipment is used for this job. It is not uncommon for the cost of casing and cement to approach half the total well cost.

Unique Aspects of Geothermal Drilling

Compared to the sedimentary formations of most oil and gas reservoirs, geothermal formations are, by definition, hot (production intervals from 160°C to above 300°C). They are often hard (240⁺ MPa compressive strength), abrasive (quartz content above 50%), highly fractured (fracture apertures of centimeters), and under-pressured. They often contain corrosive fluids, and some formation fluids have very high solids content (total dissolved solids in some Imperial Valley brines is above 250,000 ppm). These conditions mean that drilling is usually difficult – rate of penetration and bit life are typically low [5], corrosion is often a problem [6], lost circulation is frequent and severe, and most of these problems are compounded by high temperature.

Lost circulation (loss of drilling fluid into the rock formation) and reservoir damage deserve special mention. Lost circulation is often massive, with complete loss of returns at pumping rates of hundreds of barrels per hour. Geothermal wells have been abandoned because of the inability to get through a loss zone [7], and many more have needed an unplanned string of casing to seal off a problem. Lost-circulation treatment is complicated by the requirement that the treatment of loss zones must not damage the producing formation, but it is often difficult to distinguish between the two.

Finally, geothermal wells produce, relative to oil and gas, a low-value fluid – hot water or steam. For economic viability, then, geothermal flow rates and well diameters must be much larger than comparable oil and gas wells. Oil wells frequently produce through 6 cm tubing, but geothermal wells that supply power plants will generally have production intervals of at least 21.6 cm diameter. Geothermal casing will therefore be larger and more expensive and, also unlike oil and gas, it must be cemented along its complete length, not just anchored at the bottom.

All of these factors will be discussed in more detail below.

Geothermal Rock Formations

With few exceptions, geothermal reservoirs are found in igneous or metamorphic rocks such as granite, granodiorite, quartzite, basalt, and volcanic tuff [8]. Reservoirs in California's Imperial Valley and Mexico's Cerro Prieto fields are among the rare resources in sedimentary formations, and drilling practices in these fields are significantly different from elsewhere. As noted above, these igneous or metamorphic rocks tend to be hard, abrasive, and fractured, which makes drilling difficult, and they are also more variable from one well to another than is the case in a typical oil and gas reservoir. This means that the learning curve in a geothermal reservoir is not as steep as would be the case with hydrocarbons, but experience is still valuable, and each well will have a share of "lessons learned." This variability is a key factor in assessing the variables that drive well cost, as discussed in the next section.

Depth and temperature of geothermal resources vary considerably. Several power plants (e.g., Steamboat Hills, Nevada and Mammoth Lakes, California) operate on

lower temperature fluid (below 200°C) produced from depths of approximately 330 m, but wells in the Geysers produce dry steam (above 240°C) and are typically 2,500–3,000 m deep. In the most extreme cases, an exploratory well with a bottom-hole temperature of 500°C at approximately 3,350 m has been completed in Japan [9], and experimental holes into molten rock (above 980°C) have been drilled both in Hawaii and in Iceland.

Well Cost

Cost of the wells is clearly crucial to the financial viability of a proposed geothermal power project, because the well field – production and injection – can comprise 30–50% of the project's capital cost [10]. Factors that affect well cost are discussed in many places [11], and all of the topics discussed in this entry are related in some measure to the well's cost. It is useful, however, to look specifically at some of the most important cost drivers in geothermal drilling.

Well design: Design of a geothermal well is a “bottom-up” process. Location of the production zone determines the well's overall length, and the required flow rate determines diameter at the bottom of the hole – the well's profile above the production zone is then set by iteration of the successively larger casing strings required by drilling or geological considerations.

Because of the large diameters in geothermal wells, however, casing and cementing costs form a relatively large share of the cost, and the ability to eliminate one string of casing would have a major impact.

The need for directional drilling and the accuracy with which the hole trajectory must be controlled are also important factors in cost, but there is usually less flexibility in those choices as the well is designed.

Trouble: “Trouble” is a generic name for many sorts of unplanned events during drilling, ranging from minor (small amounts of lost circulation) to catastrophic (the BHA is stuck in the hole and the drill string is twisted off). In some cases, experience in the same or similar reservoirs will give a hint that certain types of trouble are likely, but at other times events are completely unexpected. It is difficult, therefore, to estimate a precise budget for trouble, but all well expenditure planning must contain some contingency funds, and this number is often taken to be around 10% of the total budget.

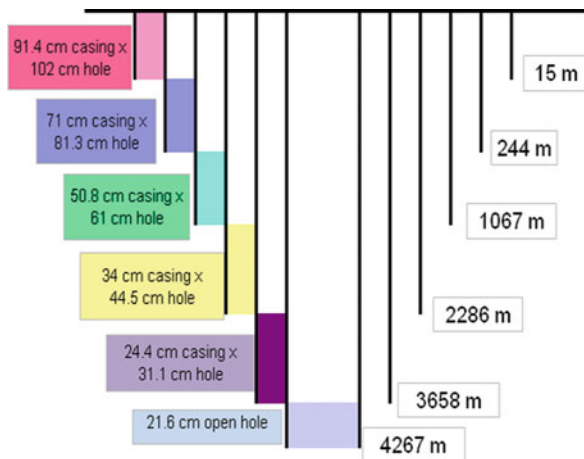
Two kinds of trouble avoidance that have received considerable research effort are lost-circulation treatment and vibration reduction to mitigate bit, BHA, and drill pipe failure. These topics are described in more detail later, but they cannot be dismissed as possible cost drivers.

Rate of penetration (ROP): Many of the costs attributed to drilling are time dependent (primarily related to the rental rate on the rig), so it is clear that anything to speed up the hole advance is beneficial. (Keep in mind, however, that increased ROP at the expense of more trips, or lower tool life, is usually not effective. See the next paragraph.) A tremendous amount of research has been done to improve bit performance, both in terms of drilling speed and life, and there is no doubt that today's bits are far better than those of an earlier generation. Still, even with improved bits it is not always easy to optimize the performance with a new bit design drilling an unfamiliar formation. The three parameters that can be easily changed for any bit/formation combination are rotary speed, weight on bit (WOB), and hydraulics (combination of jet size and flow rate), and it often takes some experimentation to determine the best combination of these factors.

Bit and tool life: Much of the commentary above about ROP applies to bit and tool life. Improved tool life means, of course, that the expense of replacing a bit or other piece of equipment can be avoided or delayed, but there is also a time saving if trips can be eliminated. This becomes more important as the hole gets deeper and the trips take more time.

Comparison: To examine the effect of these factors, the hypothetical well shown in Fig. 3 will have its overall cost calculated with the following changes in the relevant variables. Cost calculations are done with a spreadsheet program that lists and sums the major variables in drilling cost, and are generally in 2009 dollars, although the key point here is not the absolute value of well cost, but the relative effect of the various changes.

- **Well design:** The first alternative to the “base case” well is a variation designed with one fewer casing strings (this may or may not be realistic, but serves to demonstrate the effect of well design on cost) (Fig. 4).
- **Trouble:** The base case well is assumed to have a moderate amount of trouble. There are two lost-circulation events, both in the 44.5 cm interval, and



Geothermal Resources, Drilling for. Figure 3
Base case well design

one twist-off in the 31.1 cm interval. Each lost-circulation event is treated by pumping 10 m³ of cement and waiting 12 h on cement to cure, and the twist-off is assumed to require 80 h for fishing and repair.

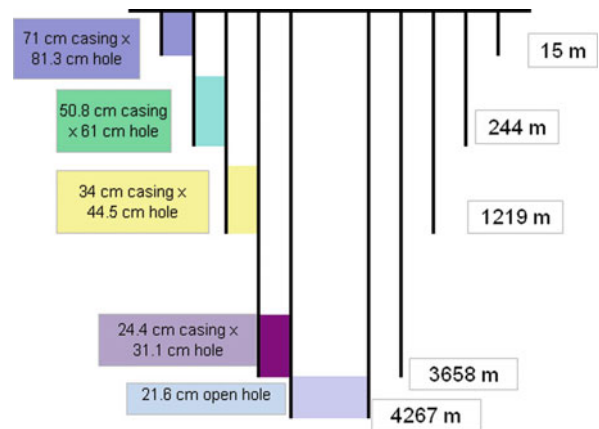
- Rate of penetration: Base case bit performance is 10 m/h for 44.5 cm bits, 5 m/h for 31.1 cm bits, and 4 m/h for 21.5 cm bits. “Improved” ROPs are 15, 10, and 6 m/h, respectively.
- Tool life: Base case bit life for the sizes above are 100, 80, and 40 h. Improved values are 200, 120, and 80 h, respectively.

Results: The effects of these improvements are shown in Table 1, where the cost savings from various changes are shown in all possible combinations.

Cost reductions that can be attributed to the various individual improvements are

- Redesign casing to eliminate one string – 18.9%
- Improve rate of penetration – 7.2%
- Improve bit life – 1.9%
- Eliminate trouble – 2.0%

From this analysis it is clear that significant cost reductions may be available. If all the postulated improvements were made, total cost would decrease by almost 29%, or more than \$3 million. Most important of the improvements considered was redesigning the casing to eliminate one string, with improved rate of penetration a distant second, although still significant. This



Geothermal Resources, Drilling for. Figure 4
Alternative casing design

statement does not imply that either of these improvements is actually possible in every case, but the numbers do give an indication of priorities when considering research into different drilling technologies. The most important factor to remember in considering well cost is that geothermal wells are *extremely* site specific, much more so than oil and gas wells of similar depth, and so these results are not generic.

The numbers above should be taken with some caution, because of this variability in well cost components with location. To give several examples of exceptions, consider the following:

- The trouble postulated for this well was relatively minor, with two cement plugs and a twist-off that could be retrieved by fishing. In at least one case [12], a geothermal well received 20 cement plugs without curing the lost-circulation problem, and the well was abandoned. As for twist-offs, it is often the case that a stuck bottom-hole assembly (BHA) cannot be fished, so that the hole must be side-tracked to go around it – this procedure is much more expensive than simply fishing.
- Differences in site preparation are not included in these spreadsheet calculations, but rugged terrain or an absence of easily available water can have a major impact on well cost.
- Some locations, such as the Imperial Valley in southern California, have extremely corrosive in situ fluids that require titanium casing [13], at a cost exceeding \$2 million for the production string.

Geothermal Resources, Drilling for. Table 1 Comparison of well costs, assuming different drilling conditions

	Base case	Improve bit life	Improve ROP	Improve ROP and bit life
Base case	\$11,560,857.95	\$11,345,822.83	\$10,733,021.31	\$10,691,820.96
Reduce trouble	\$11,335,667.78	\$11,120,632.66	\$10,507,831.14	\$10,466,630.80
Eliminate casing string	\$9,372,815.35	\$9,151,844.66	\$8,624,445.23	\$8,461,433.57
Eliminate string, reduce trouble	\$9,146,089.76	\$8,925,119.08	\$8,397,719.65	\$8,234,707.98

Geothermal Resources, Drilling for. Table 2 Sample well costs from various geothermal projects

Well location	Depth (m)	Production diameter (cm)	Year drilled	Total cost (US\$)
Newberry Caldera, USA [14]	2,927	31.1 hole/24.4 liner	1995	2,895,493
Vale OR, USA [15]	1,755	15.9 hole/12.7 slotted liner	1994	920,325
Habanero-2, S. Australia ^a	4,358	15.2 open hole	2007	6,200,000
GPK-4, Soultz, France ^a	5,260	21.6 open hole	2004	5,000,000

^aThe referenced report “The Future of Geothermal Energy,” available online at http://www1.eere.energy.gov/geothermal/future_geothermal.html, has an extensive discussion of well cost variation over periods of several decades, and also goes into much greater detail about the factors that drive well cost

- Bit-life improvements used in the spreadsheet calculations may be unrealistically high, but this only demonstrates the fact that even very large improvements in that area have relatively small effect in the given well design with the stated drilling performance. On the other hand, rock reduction is often the source of trouble cost, so the reduction in drilling time may not be the only cost saving.

In general, cost information on wells drilled by commercial geothermal operators is tightly held and it is difficult to get extensive data on actual drilling costs. In a few cases, however, such data is in the public domain, and a brief sample is given in Table 2.

Even these few examples show the great variability in cost, a significant fraction of which is caused by economic inflation over time.

Finally, if an operator is considering exploratory wells, there are many advantages to drilling “slim holes” – wells with diameters smaller than would be used for production but large enough to be useful in characterizing the reservoir. This would typically mean final diameters of 7–10 cm, compared to common production diameters of 15–24 cm.

Drilling is cheaper for slim holes than for production wells because the rigs, casing and cementing, crews, locations, and drilling fluid requirements are all smaller; because site preparation and road construction in remote areas is significantly reduced, up to and including the use of helicopter-portable rigs; and because it is not necessary to repair lost-circulation zones before drilling ahead. An extensive slimhole-drilling research program [16] showed that slim holes are consistently cheaper than full-size holes in the same locations, and that the slim holes are adequate to characterize the reservoir.

Planning and Designing the Well

There are two separate but closely related parts of preparing for a drilling project – *planning* the well and *designing* the well. “Planning” means to list, define, schedule, and budget for all the multitude of individual activities required to drill the well, and “designing” means to specify all the physical parameters (depth, diameter, etc.) that define the well itself. Detailed instructions on how to complete this process for even one well would need a sizable volume in itself, and so that is well beyond the scope of this entry, but the following discussion will

present a sort of checklist that specifies many of the questions that must be considered during these preparations. (The geographical location of the well can have a major impact on cost, schedule, and even well design, but that choice is a function of exploration for the resource, and so is too variable to be considered as a generic part of well planning.)

Careful planning is critical for any drilling operation. It will not only minimize cost, but will reduce the risk of injury or property damage from unexpected events. A drilling plan should list and define all the activities required to complete the well, with their related costs and times, and should give sufficient descriptions of individual tasks to make clear the sequence in which they must be performed. (A “critical path” approach, showing which operations must be sequential and which can be simultaneous, is often useful. The crux of this technique is that any delay along the chain of sequential operations – the critical path – will cause a delay in project completion, while delay in some other operation may not.) It is also essential that all the contractors and service companies should meet, or at least thoroughly communicate, during the planning stage, so that the plan assigns responsibilities for the various activities and there is no confusion as to what person or company performs each step.

Descriptions in the plan must be relatively detailed. For example, to specify drilling, an interval between two given depths and running casing in it would typically require, at minimum, the following information:

- Bit size and type (include suggested weight on bit and rotary speed, if available)
 - Definition of all components of the bottom-hole assembly, and whether downhole motors are to be used
 - Expected rate of penetration and bit life (thus, expected time to drill the interval)
 - Any directional drilling instructions
 - Drilling fluid type and flow rate
 - Any required logging during drilling or before casing is run
 - Size, weight, and grade of casing
 - Proposed cementing program
 - Any problems expected in that interval, or special precautions to be taken
- A plan can be as simple as a written outline, in list format, of the various activities, or can be quite detailed and in active electronic format. Management software ranges from simple spreadsheets, through freeware available on the Web, to sophisticated planning tools such as Microsoft Project [17]. If one considers commercial planning software specific to drilling, make sure that it can include services that are common in geothermal drilling but not often used in oil and gas, such as mud coolers, high-temperature tools and cement, etc. Clearly, the drilling plan must also be flexible enough to accommodate unexpected events, or trouble, during the project, and there must be a well-defined process to identify the person who is responsible for changes in the plan.
- To begin designing the well, a great variety of information is desirable, but it is not always possible to get the complete package. It is worth considerable effort to get as much of it as possible, but sometimes the designer must just go with the best available data. The desirable information includes, but is not limited to, the following parameters.
- Purpose of the well: A given well may serve any one of several different functions – production, injection, exploration, or workover – and the well design will be influenced by its purpose. For example, an exploration well might be of smaller diameter than the one intended for production and, because it might be scheduled for abandonment once the reservoir is characterized, it might also be completed with less attention to the well’s longevity (different cement, casing material, or the like). Some considerations for hole diameter in small exploration wells or “slimholes” are described in the section on [“Drilling System Selection Criteria”](#).
 - Reservoir conditions: It is extremely useful to know as much as possible about the prospective reservoir; such information might come from previous temperature and pressure logs in offset wells, nearby thermal gradient holes, or geophysical information. Clearly, temperature and pressure are crucial, but brine chemistry is also very important because it can have a major impact on casing selection and cost.
 - Logistical requirements: It is common that, for reasons including a power sales contract, other

financing requirements, or even weather, a drilling project must be completed on a given schedule. If this is the case, it can complicate planning because of factors ranging from drill rig availability to acquisition of the necessary permits. It is also more or less a standard condition that any lease site will have regulatory stipulations that affect drilling fluid disposal, cuttings disposal, possibly water supply, and even air-quality requirements that will necessitate emissions control on the rig engines. The well planner has little recourse in dealing with these factors, but it is certainly essential to consider them in the planning process.

- Likely problems in drilling: Experience in similar wells or general knowledge of the reservoir can sometimes offer a prediction of what problems may be encountered in drilling the well. If this knowledge is available, it will guide the preparations in many ways: having lost-circulation material (LCM) for under-pressured formations; appropriate drilling fluid additives for corrosive brines or for exceptionally high temperatures; high-temperature logging or steering tools and drilling motors if those tools will be used in a hot hole; and stand-by fishing tools and possibly shock absorbers in the BHA if there is likely to be rough drilling with twist-offs. It may also provide better definition of the best operating envelope (weight on bit, rotary speed, and hydraulics) for the bit in specific formations.
- Casing requirements: The heart of well design is the specification of the casing program, which will be discussed in more detail below. Parameters that determine the casing requirements include the following: nominal production rate from the well and the casing diameter implied by that flow rate, depth of the production zone, expected temperature, brine chemistry, whether the completion will be open hole or slotted liner, well trajectory (vertical, directional, or multi-leg), kick-off point (if directional), need for special casing connections, and the length of individual casing intervals.

In general, the well is designed from the bottom up, that is, the expected depth of the production zone and the expected flow rate will determine the wellbore geometry and casing program and most of the equipment requirements will follow from those criteria.

Because geothermal wells produce a relatively low-value fluid – hot water or steam – flow rates must be much higher (often $>100,000$ kg/h) than for oil and gas wells, and geothermal wells produce directly from the reservoir into the casing, instead of through the production tubing inside casing as in most oil wells. If there is two-phase flow in the wellbore, larger casing diameter where flow is vapor dominated will significantly reduce pressure drop, improving productivity [18]. Finally, many lower-temperature geothermal wells are not self-energized and must be pumped, either with line-shaft pumps driven from the surface or with downhole submersible pumps (and so the well's design must allow for pump removal). All these factors combine to drive geothermal casing diameters much larger than oil and gas wells of comparable depth – typical casing sizes in geothermal production zones are 20–34 cm.

There are three important considerations in designing the casing:

- Because each casing string limits the diameter of the drill bit and successive casing strings that can pass through it, the hole diameter decreases as the well gets deeper.
- Because of casing costs and diameter reduction, it is beneficial to make the intervals between casing points as long as possible.
- The incidence of problems or trouble increases as the wellbore intervals between casing points grow longer.

The two latter points counter each other – it is highly desirable to drill long intervals between running successive casings, but doing so greatly increases the probability of trouble. If a “contingency string” is needed to isolate a troublesome wellbore zone, this imposes a significant cost for the additional casing and cementing. It also implies the necessity of starting with larger diameter casing above the contingency string, to preserve the required bottom-hole diameter, or of completing the well with smaller bottom-hole diameter than was desired, if no provision for the contingency string was included in the plan.

Given a bottom-hole depth and diameter, determination of the casing intervals above that depends on several factors, including rock properties, formation fluids, or even regulatory requirements (some agencies

require that at least 10% of the wellbore always be behind surface casing down to the next casing point, and 1/3 the well behind casing below that). There are many common reasons to set casing in a particular interval:

- Protect an aquifer – regulations require sealing off aquifers to prevent their contamination by wellbore or drilling fluids.
- Isolate troublesome formations – these can be unstable (sloughing, swelling, or unconsolidated) formations, zones with high or incurable lost circulation, or a depleted-pressure zone above the production horizon.
- Fluid pressure control – although more common in oil and gas than in geothermal, drilling fluids often contain additives that bring the specific gravity of the fluid well above that of water, so that the weight of the fluid column will control the downhole pore pressure in the formation. This often leads to the situation in which the higher pressure of the drilling fluid exceeds the fracture gradient of the formation, leading to lost circulation or even loss of well control.
- Define the production zone – geothermal reservoirs can have more than one productive zone and casing is sometimes set to preferentially allow production from the selected zone.

There are many other reasons that casing might be set at a particular depth, but this list gives a flavor of how variable those reasons can be. Once the general casing profile is selected, the casing for each individual interval, or string, is characterized by three basic measurements: diameter, weight, and grade. Diameter is straightforward; it is just the nominal outside diameter for that interval (although this does not include the couplings, which are larger than the casing body and control the smallest possible inside diameter of the next larger string). Weight, expressed in weight units per unit length, is actually a measure of the wall thickness of the casing; heavier casing has smaller inside diameter, since the outside diameter must remain constant for a nominal size. The casing's grade is primarily related to the material's tensile strength, although there are some metallurgical variations aimed to withstand specific effects, such as corrosion, of the wellbore fluid chemistries.

Casing has to withstand different kinds of loading in different situations, and the most common design criteria are for burst pressure, collapse pressure, axial tension, and buckling. Burst pressure and axial tensile strength are a function of the casing grade, but collapse and buckling are more related to the wall thickness, because they are determined by the material's elastic properties as well as its tensile strength.

Although reasonably simple casing designs can be done with hand calculations and manufacturers' handbooks, the general topic can be very complex, and detailed procedures for casing design are well beyond the scope of this entry. Extensive resources are available. All drilling engineering textbooks [19] have sections on casing design, and an Internet search for "casing design software" will indicate the multitude of options to be found among drilling service companies. Although all of these methods are likely to produce satisfactory casing designs, engineering judgment is still important and it is of significant benefit to have a veteran drilling engineer with geothermal experience to at least review a proposed casing program.

Drilling System Selection Criteria

Most of the criteria used to select a drill rig will be derived from well parameters; specifically diameter, depth, and casing design. The process of planning and designing the well will have established the diameter, which is the primary criterion for whether the well is considered a "slimhole" or will be a conventional well and, thus, what kind of rig will be used.

Several factors define the minimum hole diameter, and also bear upon whether a core rig can be used for the hole.

- Logging tools – Typical temperature-pressure-spinner logging tools will fit into almost any reasonable hole size, but if more complex tools, especially imaging tools such as a formation micro-scanner or a borehole televiewer are to be used, the heat-shielding they require at high temperature sometimes defines a minimum hole size.
- Core size – If core is required to validate a geologic model of the reservoir or to assess the fracture dip, density, and aperture, then a coring rig is advantageous, compared with taking core samples with a rotary rig, but the core size must be considered.

Diameter is not too important for fracture data, but sometimes a rock mechanics evaluation will need a minimum core diameter. Larger diameter core also gives better recovery in highly fractured or unconsolidated formation.

- Packers – Inflatable packers are sometimes used to isolate a specific section of the wellbore for injection tests, fluid sampling, or other diagnostics. In general, this means that some kind of logging or sampling tool must be run through the packer into the zone below it, and the size of this tool will determine the minimum size of the packer and thus the hole. Based just on the diameter of the cable head for most logging cables, it would be very difficult to run a pass-through packer in a hole smaller than approximately 100 mm diameter.
- Flow test – If a flow test is expected after drilling, there are two advantages to keeping the hole diameter as large as possible: scaling up for predicted flow in a large-diameter well will be more accurate; and if the combination of depth, pressure, and temperature means that the well's ability to produce is marginal, a larger diameter hole is more likely to flow. The larger-diameter wellbore is particularly important if the flow turns two phase.

If all these factors indicate that a slimhole will satisfy the requirements, then a minerals-type coring rig can often yield significant cost savings for two reasons:

- Smaller casing, tools (bits, reamers, etc.), and cementing volumes
- The ability to drill with complete lost circulation (no returns to the surface)

Coring rigs (see Fig. 5) are fundamentally different from rotary rigs in the way that they retrieve core. A typical coring rig used for minerals exploration stores the core as it is cut in a tube in the lower end of the drill string. At the end of the coring run, a wireline is lowered down the inside of the drill string and is latched into the top of the core tube to retrieve it to the surface. This not only gives a continuous core over the interval of the hole, but is much faster than tripping the drill string to retrieve the core sample as is done in rotary rigs.

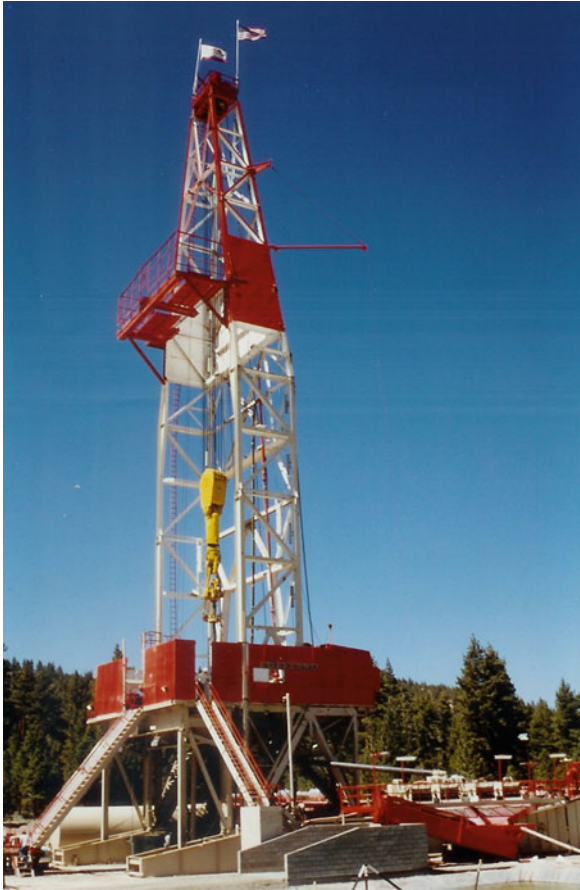
If a large-diameter hole is required, then a conventional rotary rig (see Fig. 6) will probably be



Geothermal Resources, Drilling for. Figure 5
Typical coring rig, mast is ~15 m high

used and the basic choice to be made is whether it should be a top drive. For many years, as described in the Overview, the drill string was turned by a “rotary table” in the rig floor. A square or hexagonal bushing in this table applies torque to the “kelly” (the topmost part of the drill string), which is square or hexagonal in cross section, so that it can be turned by the table and still slide downward as the hole advances.

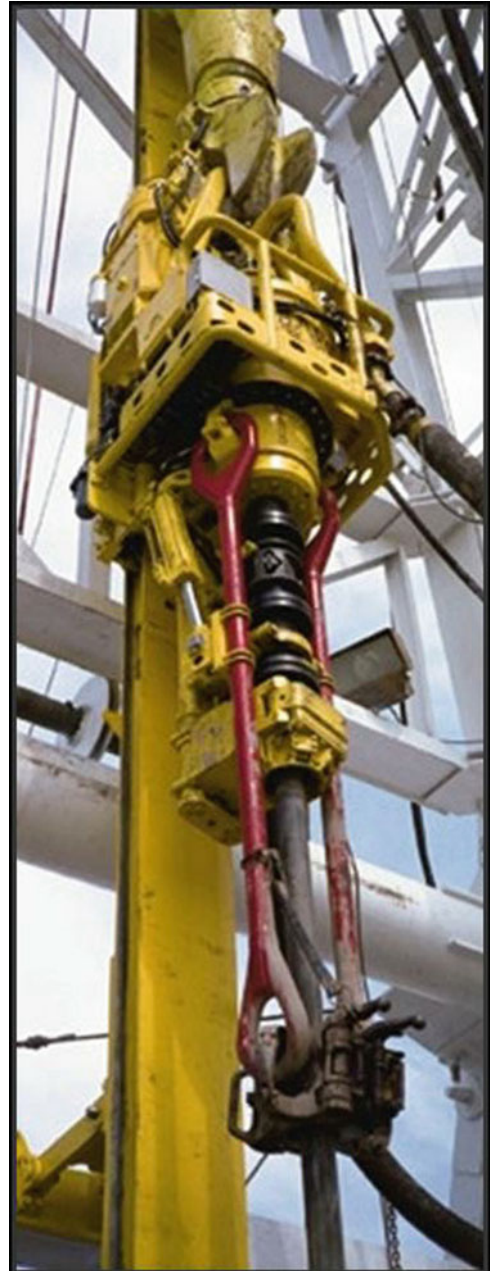
In the early 1980s, however, a new system in which the drill string was turned by a motor hanging directly beneath the traveling block gained commercial acceptance (see Fig. 7). This “top drive” technique has at least two critical advantages: instead of adding drill pipe one joint at a time as the hole advances, the driller can work with stands (two or three joints) of pipe,



Geothermal Resources, Drilling for. Figure 6
Rotary drill rig, mast is ~55 m high

eliminating time and connections, and the driller can rotate and circulate while tripping. Detailed comparison of operations for one offshore platform [20] showed an 11% decrease in drilling time, and the ability to circulate while tripping is especially important for geothermal wells, because it allows protection of temperature-sensitive tools while tripping into the hole. In one geothermal reservoir [21], it was reported that bit life was improved three- to sixfold by circulating during trips into the hole. The circulation/rotation capability is also useful for avoiding stuck pipe and for working through tight spots during tripping. Top-drive rigs generally cost more in daily rental, but it is often cost effective to use one.

Many considerations will affect the final rig choice but, aside from the purely economic factor of the price quoted by the drilling contractor, the following aspects



Geothermal Resources, Drilling for. Figure 7
Top drive, photo courtesy of National Oilwell Varco

of the rig should be the minimum list of qualities upon which to make a decision.

Rig capacity: This usually refers to hook load – the weight that can be suspended from the rig's hoisting system. Clearly, the drill string weight, with all the bottom-hole assembly, is an important part of this

requirement, but it should be remembered that the casing is often the heaviest load handled during a drilling project.

Rig footprint: The drilling contractor should provide a dimensioned diagram or map of the rig setup in operating mode. It should clearly show: access points and traffic patterns to various parts of the rig; where different operations (mixing mud, mud logging, etc.) are performed; and the locations where various consumables are stored. If the planned drilling operation includes mud pits, or a water well, those should also be on the map. The contractor's quote will give a cost figure for mobilizing and demobilizing the rig (moving the rig to and from the drilling location) but there should also be an indication of how many truck loads this will entail, and what road clearances are required, in case there are regulatory issues at sensitive locations.

Pump capacity: As discussed under section [Drilling Fluids](#), the pumps must have enough volumetric capacity to give sufficient velocity in the annulus to lift the cuttings. The pumps must also have enough pressure capacity to give the desired pressure drop through the bit jets, and possibly drive a downhole drilling motor, if that is planned or a likely contingency.

Fluid cleaning: These requirements should be defined in consultation with the mud engineer/company, and the rig's shakers, desanders, desilters, and centrifuges should be adequate to the job. There should be some operational consideration of the rig's compatibility with any environmental regulations that affect disposal of the drilling fluid and cuttings, such as the requirement for a closed-loop fluid system with no discharge to the environment.

Drill string and BHA: The bottom-hole assembly design should be defined during planning, so it is clearly important that the rigs have the correct tools, tongs, and fixtures (bit breakers, elevators, etc.) to handle all components of the drill string. It should also be made clear in the contractor's quote whether drill pipe is included in the rig's daily rate. If so, the planner should make certain that it is the correct weight and grade and, if not, the planner should assure that another source of pipe is available.

High-temperature capability: When drilling geothermal wells, it is clearly necessary that any of the rig's downhole or surface equipment that will be exposed to high temperature has that capability. This may be

especially noticeable in drilling fluid returns, which will probably be much hotter than in conventional drilling. In most locations, regulatory guidelines will require use of mud coolers when returns exceed a specific temperature, but even with coolers, operating personnel should be aware that hot fluid will create higher-than-normal thermal expansion forces, and that any elastomer seals may become vulnerable to the high temperature.

Rig instrumentation: Complete information about the rig's performance is essential for safe, efficient operation, and the project planner should include an instrumentation list in the rig criteria. Detailed requirements will vary from project to project, but a typical set of desirable measurements includes the following: drilling fluid inflow and outflow rates, drilling fluid inflow and outflow temperatures, standpipe pressure, rotary speed, weight on bit, torque, and kelly height, if available. All these measurements should be digitally recorded on a data logger at reasonably short intervals (~ every 5 s) so that they can be easily stored and retrieved, but selecting the interval between measurements is not straightforward. For "steady-state" drilling, in which operations are routine, data points every 5–10 s are adequate, but for transient events such as the beginning of a new bit run or the onset of unstable, possibly damaging, drilling conditions, high-resolution data can be extremely valuable. Collecting high-speed data implies very large data files on long drilling projects, which may be a storage problem, but low-speed collection that gives more manageable amounts of data may not give the resolution needed for short-duration events.

Rig instrumentation is often coordinated between the drilling contractor and mud-logging company; see the Instrumentation section below.

Support: In general, rig malfunction or breakdown is one of the less likely kinds of drilling trouble. If the drill site is in an especially remote location, however, it is worth considering how far away the rig's support services may be.

Crew and training: It is not always possible to know in advance who will be working on the rig, but the importance of a well-trained, experienced crew to the project's success cannot be overstated. In the course of evaluating proposals from drilling contractors, every effort should be made to find out the experience and qualifications of the rig crew and supervision.

Like many aspects of drilling, selecting a rig often turns out to be more complicated than it first appears. Keys to a successful choice revolve around having a clear and detailed concept of what is needed for the project. It is frequently very valuable to have an experienced geothermal drilling engineer assigned to the specific task of rig selection, because any extra cost incurred here will almost certainly prove to be well spent.

Drill Bits and Bottom-Hole Assembly

As described in the Overview, the bit is at the lower end of the drill string where, as it rotates, it crushes, gouges, grinds, and cuts the rock to advance the hole. The bottom-hole assembly (BHA) comprises all the components, including the bit, up to the lower end of the drill pipe.

Bits

The bit is usually either a roller cone, which crushes and gouges the rock as the cones turn and their teeth successively come in contact with unbroken areas, or a drag bit, which shears the rock in the same way that a machine tool cuts metal. Because of this shearing action, drag bits are inherently more efficient than roller-cone bits.

The great majority of roller-cone bits today have three cones, with either milled steel teeth (see Fig. 8) that are part of the cone itself or hard-metal (usually tungsten carbide) teeth (see Fig. 9) inserted into the body of the steel cone. Milled-tooth bits are less expensive but are suited only for softer formations. Insert bits are used in medium to harder formations, with the size, shape, and number of inserts varied to fit the specific drilling conditions. The bits are available with either roller or journal bearings, depending on operating conditions, and the bearings, seals, and lubricants should all be specified to withstand high temperatures if the bits are to be used in geothermal drilling. Roller-cone bit technology is very mature – over 100 years since the first patent [22]. Although bit companies still do constant research, and have made significant progress over the last 20 years, the improvements have been incremental. Since the 1950s, R&D for roller-cone bits has alternated between better bearings and more durable cutting structures, depending on which is



Geothermal Resources, Drilling for. Figure 8
Milled tooth roller-cone bit (Photo courtesy of Reed-Hycalog NOV)

the dominant failure mode at the time. Roller-cone bits dominate drilling for geothermal resources because of their durability in the hard, fractured rocks that are characteristic of those reservoirs.

Because drag bits reduce rock with a shearing action, they are inherently more efficient than roller-cone bits. Drag bits with polycrystalline-diamond-compact (PDC) cutters (see Fig. 10) began to be widely used in the early 1980s for their ability to drill faster and last longer in soft to medium formations, and they now dominate oil and gas drilling. A particular advantage of drag bits for geothermal drilling is that they do not have any moving parts, so temperature limitations on bearings, seals, and lubricants are not a factor. Unfortunately, PDC bits usually do not have acceptable life in hard or fractured formations, and are not generally used in geothermal drilling. A great deal of work has been done to extend their use to harder rocks [23, 24, 25], and significant progress has been made, but acceptance by the geothermal industry has been minimal. Wider use of these more efficient bits would be a significant technology advance.



Geothermal Resources, Drilling for. Figure 9
Insert roller-cone bit (Photo courtesy of Reed-Hycalog NOV)



Geothermal Resources, Drilling for. Figure 10
PDC drag bit (Photo courtesy of Reed-Hycalog NOV)

If the drilling plan calls for a slimhole to be drilled with a minerals-type core rig, bits are completely different. Much of the rock volume removed from the hole is in the form of core, and the rock cuttings themselves are much smaller, because virtually all hard-rock coring is done with diamond-impregnated bits (see Fig. 11) that grind away the rock.

Principal variations in this kind of bit are the diamond grain size, the diamond grain density, and the hardness of the matrix metal in which the diamond grains are embedded. These bits typically turn at much higher speeds than conventional rotary bits (either roller cone or drag) and have a much lower drilling fluid flow rate because of the smaller annulus between the drill rods and the borehole wall.

Drill Pipe

Choosing the drill pipe specifications can be complicated in some cases, but the primary considerations are the following.

- **Strength:** The principal requirements are for tensile and torsional strength, so that the pipe can pull the drill string out of the hole (often with some overpull required because of tight spots, or even partially stuck pipe) and can apply the torque needed to rotate the bit. Internal pressure may become an



Geothermal Resources, Drilling for. Figure 11
Diamond-impregnated core bit

issue in some cases, and bending strength is important in directional drilling.

- **Size:** Given that several different pipe configurations might be strong enough, a major driver for size selection is hydraulics. The internal diameter of the pipe must be large enough to avoid excessive pressure drop in the circulating drilling fluid. It is also necessary that the inside diameter of the pipe be large enough to pass any expected logging tools, and there are sometimes considerations of whether the pipe size is adaptable to fishing tools in the event of trouble. On the other hand, the outside diameter of the drill pipe tool joints must clearly be small enough to pass through the smallest casing to be used, with enough clearance for the same fluid flow on the outside of the pipe.
- **Corrosion resistance:** Many formation fluids are corrosive; this is especially true in much geothermal drilling. Several special grades of drill pipe are made from alloys designed for corrosive environments.
- **Wear resistance:** Because many geothermal formations are extremely abrasive, drill pipe tends to wear much faster than in other types of drilling. “Hard banding” (applying layers of wear-resistant material such as tungsten carbide to the outside diameters of the tool joints) is common in geothermal drilling.

Bottom-Hole Assembly (BHA)

A drill string is relatively flexible compared to its length (a scale model, dimensionally, of a 3,000 m drill string is a piece of steel wire, the thickness of a human hair, one meter long). The total weight of the drill string is generally much greater than the desirable force on the bit, so the rig’s hoisting capability holds back some of the string weight to control force on the bit. The upper part of the drill string is therefore in tension, while the lower part that applies force to the bit is in compression. Drilling with the relatively thin drill pipe in compression may cause buckling, so it is important that the neutral point (where the drill string stress changes from tensile to compressive) falls within the *drill collars*, which are thick-walled cylinders at the bottom of the drill string. The outside diameter of the collars is controlled by the necessary annulus between the collars and the wellbore, the inside diameter by hydraulic consideration (large enough to prevent excessive pressure drop), and the

overall length by that required to provide maximum expected weight on bit. Other components that are often part of the BHA include the following.

- **Stabilizers:** Because the drill collars and other components must be smaller than the wellbore diameter to provide a path for fluid circulation, they can have major lateral deflections. This can produce serious vibration as well as high fatigue loads in the threaded connections, so stabilizers that have full wellbore diameter on ribs along the outside surface but leave a flow path between the ribs are widely used at multiple points in the bottom-hole assembly.
- **Reamers:** The outside diameter, or “gauge” of drill bits tends to wear, causing the hole to be smaller than the nominal diameter. When a new bit is tripped in, it has to ream the smaller hole out to the desired diameter, which is time consuming and which causes the new bit to wear prematurely on its own outside diameter. Additional cutting elements, either as fixed cutters or as toothed, cylindrical rollers are often added to the BHA just above the bit, to help maintain the full hole diameter.
- **Shock absorbers:** When drilling in hard or fractured formations, or those in which soft and hard stringers are interbedded, high vibration loads are common. Shock absorbers, or dampers, are used to attenuate the vibrations transferred to the upper part of the BHA and drill string.
- **Jars:** If the drill string is stuck in the hole, it can sometimes be released by the impact force produced by jars. These function by suddenly releasing energy stored in the drill string by pulling up on it and stretching it. The two principal types are mechanical jars and hydraulic jars, but both operate on the same principle. Jars are generally used when fishing, but some drillers prefer to have jars already in the drill string during normal drilling.

Directional Drilling

During normal drilling, the pendulum effect of the heavy drill collars tends to keep the hole vertical, but for many of the following reasons it is often necessary to guide or steer the hole’s trajectory in a specific direction – institutional, legal, or topographic issues prevent the drill rig from being directly over the geologic target; it is economical

to drill several wells from one prepared site; and, particularly for geothermal wells, it is important for the wellbore to intersect as many formation fractures as possible.

Directional drilling is a relatively complex technology and there are a number of ways to drill a deviated hole, but the most common is to use a downhole motor (hydraulically powered by drilling fluid flowing through it) that turns the drill bit without rotating the drill string. A “bent sub” points the motor and bit at a slight angle to the axis of the drill string or a “bent housing” introduces an angle between the motor and the bit, and since there is no rotation, the bit continues to drill in the direction it is pointed. The difficulties inherent in directional drilling are aggravated in geothermal wells because both the electronic tools used to control and survey the well trajectory and elastomer elements in the motors are susceptible to high temperature. Progress has been made in both of these areas, but it is still often a technical challenge.

Drilling Fluids

Overview

Drilling fluid flows down the drill pipe, through nozzles in the bit, and back up the annulus between the pipe and wellbore wall, carrying the cuttings produced by the bit’s action on the formation. (An alternative method, called reverse circulation [26], is sometimes used – the fluid flows in the opposite direction, down the annulus and up the inside of the drill pipe, but it is not common – see the Section [Lost Circulation](#).) Drilling fluids can be either liquid or gas, and liquid-based fluid is universally called “mud” because the first fluids were just a mixture of water and clay.

Drilling mud is made up of three principal components:

- **Base liquid:** Oil, freshwater, or saltwater can be used as a base liquid in drilling muds, but oil and saltwater are almost totally restricted to hydrocarbon drilling. Freshwater muds are used for geothermal drilling.
- **Active solids:** Active solids are the clays and polymers added to the water to produce a colloidal suspension. They determine the viscosity of the mud and are known as viscosifiers.
- **Inert solids:** Inert solids are those added to the mud either by drilling (i.e., particles of the formation) or by using barite as a weighting material. These solids increase the density of the mud without appreciably affecting the viscosity.

Historically, most geothermal drilling fluids have been a fairly simple mixture of freshwater and bentonite clay, possibly with polymer additives [27]. Air drilling is relatively common, especially in areas like the Geysers in northern California, where the reservoir produces dry steam, and air drilling also has advantages in drilling performance because the rate of penetration is usually higher with lighter fluid. Aerated mud has a gas, usually air but sometimes nitrogen if corrosion is serious, injected into it to lighten it, and is also common where lost circulation is a significant problem.

Drilling Fluid Functions

As noted above, the principal function of drilling fluid is to clean the hole of cuttings, but there are several other purposes:

- **Cool and clean the bit:** keeping the bit cool, especially if it has elastomer seals, is critical to its life.
- **Lubricate the drill string:** this can be a significant factor in deviated (non-vertical) wells, where the drilling string is lying against the wellbore wall.
- **Maintain the stability of the borehole:** the proper drilling fluid can help control swelling or sloughing formations, thus lessening the risk of stuck drill pipe. It is also important that the fluid hold the cuttings in suspension when circulation is stopped, so that they do not fall back and pack around the bit and BHA.
- **Allow collection of geological information:** the cuttings brought back to the surface by the fluid help to identify the formation being drilled.
- **Form a semipermeable filter cake to seal the pore spaces in the formations penetrated;** this prevents fluid loss from the wellbore.
- **Control formation pressures:** if high downhole pressures are present or expected, dense material can be added to the drilling fluid to increase its specific gravity, thus resisting the downhole pressure.
- **Transmit hydraulic horsepower:** this power can be used for driving a drilling motor or for cleaning the hole and/or the bit.

Drilling Fluid System

It should be emphasized that the drilling fluid is part of a *circulating system*, comprising the fluid itself, the mud pumps, and mud-cleaning equipment. The pumps must have sufficient capacity (flow rate and pressure) to provide adequate bottom-hole cleaning, high annular velocity to lift the cuttings, and enough hydraulic horsepower to drive downhole motors and provide the designed pressure drop through the bit jets.

When the cuttings-laden mud returns to the surface, it passes through a series of devices to remove the cuttings. The first of these is usually the shale shakers, which have tilted, vibrating screens that filter out larger cuttings and let them slide off into collection containers; next are usually hydrocyclones, which use fluid inertia to swirl the fluid in a conical chamber, letting the solids drop out the bottom; and finally, centrifuges spin the fluid to extract the finest particles through their density difference. Effective mud cleaning is important for drilling performance as well as cost control. If the fluid has to be discarded because of inadequate cleaning, it is expensive both in material cost and in time loss.

Drilling Fluid Properties

The drilling fluid will be designed to have certain properties, and it is critical to monitor and control these properties at all times. Design and maintenance of drilling fluids is a complex topic, covered in great detail in many sources [28, 29] but primary attributes of fluid for a given well include the following.

- Viscosity: it is vital that the fluid's viscosity be high enough to lift cuttings out of the well as the fluid circulates, and to hold the cuttings more or less in suspension when circulation is stopped.
- Density (or specific gravity): if formation pressures are expected to be high, then the fluid can be weighted to help control them but, as is often the case in geothermal wells, if formation pressures are low, then the fluid should be as light as possible to avoid lost circulation.
- pH: the alkalinity of the fluid is important for corrosion control and for its reaction with certain formation constituents; normal pH is 9.5–10.5, but higher values are not uncommon.

- Filter cake: this is a measure of how well the fluid forms an impermeable layer on the borehole wall to prevent leakage into the formation's natural permeability. (This is typically more important in oil and gas drilling than in geothermal.)
- Solids content: this is a measure of how well the mud is being cleaned, and can also determine when the mud should be discarded or diluted.

There are standard procedures [30] for testing these and other parameters of the drilling fluid, and this testing is normally done at least daily in the field by the drilling fluid specialist or “mud engineer.”

Successful mud systems need at least these three attributes:

- Stability: The desired properties of the fluid, once established, should be stable under normal drilling conditions.
- Easy treatment: If the desired fluid properties are lost, treatment should be available to restore them.
- Property testing: Tests and testing equipment should be available to identify fluid properties and indicate any treatment required.

Although the underlying principles of drilling fluids described in the extensive literature are the same for oil/gas and geothermal drilling, high temperatures affect many of the clays and additives used to tailor the fluid properties. Some considerations unique to geothermal drilling are listed below, based heavily on the cited reference [31]:

- Viscosity control: high-quality bentonite clay is the principal viscosifier used in geothermal drilling. Several polymers, available both in liquid and powder form, are also useful but they tend to degrade at high temperatures over long periods of time, so their principal use is for high-viscosity sweeps to clean the hole before cementing, trips, or other activities that require stopping circulation. It is also sometimes necessary to decrease the viscosity, if drilled solids or high-temperature gelation have raised it too high. Proprietary blends of low-molecular-weight polymers and starch derivatives have recently been developed and are effective both in thinning the mud and in inhibiting gelation.
- Solids removal: at high temperatures, the drilled solids tend to take up the available water more

vigorously than at lower temperatures, so effective mud cleaning is even more important than usual to prevent gelation and viscosity increase.

- Filtrate (water loss) control: in the past, geothermal filtrate requirements were often more rigorous than necessary. It is important to analyze the filtrate requirements, not only for each well, but for each interval, so that expensive additives are not used without good cause. Lignite has long been the most common geothermal water-loss reducer, but proprietary polymers are also becoming common.
- Alkalinity: high pH is necessary to control the effect of some wellbore contaminants (CO_2 and H_2S), to reduce corrosion, and to increase the solubility of some mud components (lignite, etc.). Addition of caustic soda (NaOH) has been the traditional method of increasing alkalinity, but caustic potash (KOH) is becoming more common in geothermal drilling because of its benefits to wellbore stability.
- Lubricity: the drill string sometimes needs extra lubrication when directional drilling, and lubrication is very often needed when core drilling, especially when drilling without returns. Hydrocarbon-based lubricants often lose their effectiveness at high temperature, but there are proprietary, environmentally friendly lubricants that offer good performance at sustained high temperature.

Finally, there are instances in which it is desirable to drill with clear water, or clay-free drilling fluids, especially in production zones where conventional clay-based muds create a risk of formation damage. This technique requires a copious water supply, and cannot be used in all wells, but has proven successful in Iceland and in Mexico [32].

Planning the Mud Program

Some general guidelines [33] for planning the drilling fluids program are given below, with a reminder that every well is different and there are very few, if any, generic procedures that can be used without modification. A pre-spud meeting of all operating, drilling, environmental, and service company personnel is highly recommended. Discussions of the drilling plan and contingencies may eliminate trouble later in the program. Once there is agreement on the drilling plan,

then the mud program should be planned with the following considerations.

1. Water: Since water is basic to the mud system, it is important to know the quality, quantity, and cost involved with the makeup water. Poor quality makeup water may require chemical treatment prior to its use.
2. Type and thickness of the geologic strata: This is not always known before drilling, but fluid properties must be planned with the best available information about downhole conditions, that is, the reactions between drilling mud and formation.
3. Site accessibility: Make sure that supply trucks have reasonable access to the site and that rig placement in relation to pits, bulk storage, etc., is convenient to reduce handling.
4. Climate: Extremes of heat, cold, and precipitation can affect the mud system and products.
5. Drilling equipment: Make sure that the surface equipment, such as pumps, mixing and circulating tanks, mixing equipment, and solids control capabilities are adequate for the hole size, downhole tools, etc.
6. Environmental considerations: If at all possible, use nontoxic, easily disposed drilling fluids. All personnel should know all regulations pertaining to the job.
7. Manpower: The experience, skill, supervision, and attitude of the rig crews are of paramount importance to a successful drilling program.

This chapter is intended to give some flavor of the complexity of the process that is designing and maintaining a drilling fluid system. It is worth a great deal of attention in preparation for a project, because a high percentage of the problems encountered in drilling are related in some way to the fluids.

Lost Circulation

The most expensive problem routinely encountered in geothermal drilling is lost circulation, which is the loss of drilling fluid to pores or fractures in the rock formations being drilled. In addition to the cost of the drilling fluid itself, the fluid loss and inadequate hole cleaning can create many other drilling problems including stuck drill pipe, damaged bits, slow penetration rates, and collapsed boreholes. Lost circulation represents an average of 10% of total well costs in

mature geothermal areas [34] and often accounts for more than 20% of the costs in exploratory wells and developing fields. Well costs, in turn, represent 35–50% of the total capital costs of a typical geothermal project; therefore, roughly 3.5–10% of the total costs of a geothermal project can be attributable to lost circulation.

Combating lost circulation can be approached in different ways – drill ahead with lost circulation; drill with a lightweight drilling fluid that will have a static head less than the pore pressure in the formation; mix the drilling fluid with fibrous material or particles that will plug the loss apertures in the formation; or pause in the drilling and try to seal the loss zones with some material that can be drilled out as the hole advances.

Drill with Lost Circulation

Under some conditions, it is practical to drill without returns, particularly in the case of core drilling, where the cuttings are very fine and where most of the rock comes out of the hole in the form of core. In many slimhole exploration holes, intervals of hundreds of meters have been drilled with complete lost circulation [16], but this can only be done if the formations are competent enough to remain stable.

Another technology that is useful with lost circulation is dual-tube reverse circulation [35] (DTRC). This method uses a drill string of two concentric tubes, with the drilling fluid passing down the annulus between the inner and outer tubes, circulating out through the bit, and carrying the cuttings back up through the center tube. This means that it is only necessary to maintain fluid around the bit and bottom-hole assembly, so drilling with complete lost circulation is possible. This technique has been used on several geothermal wells [36] and in one case [37] reduced the cost per foot of drilling comparable wells by more than one-third.

Lightweight Fluids

Aerated fluids – liquid with gases injected into it – produce a static head less than the pore pressure and are a common remedy for lost circulation in geothermal drilling. Aqueous (water-based) foam is attractive because of its simplicity, but it is important to use the proper surfactant that has stable properties at high temperature. Considerable modeling was done in the

early development of aqueous foam for geothermal drilling [38, 39]. In addition to numerical models of the foam structure and rheology, a laboratory flow loop measured pressure, temperature, and flow rate at different points to allow experimental confirmation of a rheological model.

Aerated drilling is now used extensively in many locations, and recent experience has shown that its use not only avoids problems with lost circulation but may improve the well's productivity after drilling [40].

Lost-Circulation Materials (LCM)

Lost-circulation problems can generally be divided into two regimes, differentiated by whether the fracture aperture is smaller or larger than the bit's nozzle diameter. Clearly, LCM particles that will plug the bit are unacceptable, but for smaller fractures or for matrix permeability, the wellbore can theoretically be sealed by pumping solid or fibrous plugging material mixed with the drilling fluid – this method is much less effective with larger fractures. Very many substances have been used in the oil and gas industry to plug lost-circulation (LC) zones, but most of them have been organic or cellulosic materials that cannot withstand geothermal temperatures. LC zones in oil and gas also tend to be dominated by matrix permeability rather than the much larger fracture apertures common in geothermal reservoirs. Although traditional organic LCM can be used in the upper, cooler, intervals of a well, and several candidate materials that will withstand high temperature have been identified [41], LCM, in general, has often been unsuccessful in geothermal drilling.

Wellbore Sealing

Fractures too large to be plugged by LCM can only be sealed by withdrawing the drill string from the hole and injecting some liquid or viscous material that will enter the fractures, solidify to seal them, and then have its residue removed by resumption of drilling. Conventional lost-circulation treatment practice in geothermal drilling is to position the lower end of an open-end drill pipe (OEDP) near the suspected loss zone and pump a given quantity of cement (typically 10 m³) downhole. The objective is to emplace enough cement into the loss zone to seal it; however, this does not always occur. Because of its higher density relative to the

wellbore fluid, the cement often channels through the wellbore fluid and settles to the bottom of the wellbore (the larger diameters of geothermal wells aggravate this problem, compared to oil and gas). If the loss zone is not on bottom, the entire wellbore below the loss zone must sometimes be filled with cement before a significant volume of cement flows into the loss zone. Consequently, a large volume of hardened cement must often be drilled to reopen the hole, which wastes time and contaminates the drilling mud with cement fines. Furthermore, because of the relatively small aperture of many loss-zone fractures, the loss zone *may* preferentially accept wellbore fluids, instead of the more viscous cement, into the fractures. This causes dilution of the cement in the loss zone and loss of integrity of the subsequent cement plug. As a result, multiple cement treatments are often required to plug a single loss zone, with each plug incurring significant time and material costs. At least three different approaches have tried to improve this process.

- **Cementitious mud:** As implied by the name, this is drilling fluid with cement and other materials added to satisfy the criteria: (1) compressive strength above 3.4 MPa after 2 h cure, (2) permeability to water < 10 millidarcies, and (3) volume increase with curing. Brookhaven National Laboratory found that rapid-setting, temperature-driven cement could be formulated by mixing conventional bentonite mud with ammonium polyphosphate, borax, and magnesium oxide [42]. Significant compressive strength was developed by such admixtures in less than 2 h when sufficient concentrations of the magnesium oxide accelerator were used, and the setting time decreased with increased temperature. Furthermore, the material expanded approximately 15% upon setting. These setting characteristics were ideal for plugging major-fracture loss zones, but more control over the setting process was necessary to ensure that the cement would not set up inside drill pipe during field application.
- **Better cement placement:** Sandia National Laboratories developed a drillable straddle packer (DSP) [43] as a way to improve the effectiveness and reduce the cost of a typical cement treatment by controlling the cement flow into the loss zone and

by reducing dilution of the cement caused by other wellbore fluids flowing into the loss zone. An assembly on the end of the drill string carries two fabric bags that straddle the loss zone and provide zonal isolation. The bags are inflated with cement and seal against the wellbore wall, thereby forcing most of the cement to flow into the loss zone. After pumping a specified volume of cement, the straddle packer assembly is disconnected from the drill string and left in the wellbore while the drill string is tripped out of the hole. The packer assembly is constructed of drillable materials: aluminum, fiberglass, and, in some applications, CPVC plastic – after the cement sets, the DSP is drilled out and the operation resumes. This device was successfully tested in a full-scale wellbore and complete design drawings are contained in the reference, but it was never commercialized.

- **Polymeric grout:** The concept of using polyurethane grout instead of cement to seal fractures was investigated in the 1980s but early efforts were not successful [44]. Recent encouraging laboratory work and the growing use of polyurethane grouting in civil engineering projects [45], however, stimulated new interest in this technology. An opportunity to evaluate polyurethane grout in the field came with a DOE grant to Mt. Wheeler Power that required reopening a well near Rye Patch NV. This well had been temporarily abandoned after 20 cement plugs had failed to cure lost-circulation problems, but a prototype grouting apparatus, combined with DTRC, was successful in sealing a loss zone approximately 6 m in length and allowing the well to be reopened [46]. The polyurethane grout used in the Rye Patch well is not suited for higher geothermal temperatures, but other polymeric grouts have been developed [47] that can withstand 260°C for 8 weeks.

Despite the demonstration of methods described above, familiarity with cementing practice and ready availability of the equipment and materials mean that it is still the dominant method of formation sealing today.

Well Control

Well control, in general, has to do with controlling the flow of drilling fluids and formation fluids out of the

wellbore. If the hole advances into a fractured or permeable stratum where the pore pressure is higher than the static head of the drilling fluid, the formation fluid will flow into the wellbore – this is called a “kick” – and that flow must be controlled. If control of that flow is lost, then the resulting disaster is a “blowout” which, at the least will be very expensive and, at worst, can result in loss of life, equipment, and the drill rig.

The primary method of detecting a kick is to compare measurements of the drilling fluid inflow and outflow; if outflow is greater, there is a kick, if inflow is greater, there is lost circulation. Traditional methods of measuring these flows have been a stroke counter on the mud pumps (a volumetric calculation gives fluid inflow) and a paddle meter on the return line (a flat vane extends into the mud returns such that the angular displacement of the paddle indicates flow rate). Each of these techniques has inaccuracies: pump efficiency (and therefore displacement per stroke) varies with wear and clearances on the pumps, and the paddle meter can be influenced by any number of variables [48]. Development has shown that better methods (magnetic or Doppler flow meters for inflow and rolling float meter for returns) are available [49], and if well control is expected to be an issue, these methods should be investigated.

The apparatus that controls a kick and potential outflow at the wellhead is called the blowout preventer (BOP) or blowout prevention equipment (BOPE). The BOP stack comprises three types of device to shut off the wellbore and prevent fluid flow out of it: annular preventers, pipe rams, and blind rams. The basic function of each is to shut off the wellbore, but they operate in slightly different ways.

- Annular preventer: This is an inflatable bladder that seals around drill pipe, casing, drill collars, or irregularly shaped components of the drill string. It usually has the lowest pressure and temperature ratings of the stack components.
- Pipe rams: These are two sliding gates, each with a semicircular cutout, that come together from each side of the drill pipe. The hole in the center fits and seals around the outside diameter of the drill pipe.
- Blind rams: These are also sliding gates, but there is no hole in the center; they are used when the drill pipe is out of the hole.

Below the BOP stack, two-valved lines (called the choke and kill lines) are connected to the wellhead so that fluids can be either released from or pumped into the wellbore as part of the well-control process. There will usually be detailed regulatory requirements for the BOPE (see the California manual [50], for example, which is also an excellent reference for information on BOPE) but the critical factors are to make sure that the BOP pressure rating is adequate and that all the elastomer seals in the equipment are qualified for high temperature. One of the primary well-control techniques for geothermal drilling is simply to pump cold water down the well, so it is also important to make sure an adequate water supply is available.

In contrast to oil and gas wells, which are often overpressured and where those pressures are controlled by weighted drilling fluids, geothermal wells most often are under-pressured. This means that the formation pressure is *less* than the drilling fluid head, which is the effect that causes lost circulation, as discussed in a previous section. There are exceptions such as wells in Cooper Basin, South Australia, with wellhead pressures of approximately 35 MPa [51] but the principal issues in geothermal well control usually involve unexpected steam flow. This can be caused by drilling into a formation that is at much higher temperature, and circulating the hot fluid up the wellbore, or much higher pressure than predicted, such as an event that occurred in Hawaii [52], or by sudden, major lost circulation. Any of these events, can drop the drilling fluid level to the point that its static head no longer exceeds the saturation pressure at the formation's temperature, and either the drilling fluid or formation fluids flash into steam. Unexpected steam flow in permeable formation that is not completely sealed by casing is particularly dangerous, because steam can begin to flow up the outside of the previous casing string, eventually destroying the casing's integrity and often causing loss of the drill rig [53].

As in many contexts, prevention of a problem is more efficient than a cure. A number of methods are available to estimate the wellbore temperature profile and warn that a problem may be near: comparison of drilling fluid inflow and outflow temperatures, maximum-reading thermometers either run just above the bit or lowered through the drill pipe on a wireline, or onboard logging tools that can transmit temperature

data in real time. Although none of these is guaranteed to provide early warning of a potential kick, it is always important to know as much as possible about the downhole environment.

Having discussed above the problems of steam flow in the wellbore, however, it should be noted that in reservoirs with a dry (superheated) steam resource, such as the Geysers, the production interval is drilled with air to avoid formation damage and plugging [54]. This means that the drilling returns include produced steam from the reservoir. The top of the wellbore is closed by a “rotating head” that seals around the drill pipe, while allowing it to rotate and move downward. The gaseous returns are sent through a manifold called a “banjo box” below the BOP but above a wellhead valve, and then to the “blooie line,” which exhausts a distance away from the drill rig and where the returns receive chemical treatment for H₂S abatement. This is very similar to the technique called “managed pressure drilling” in oil and gas reservoirs, where it has been discovered that productivity is much improved if drilling fluid has not been forced into the formation by excessive downhole pressure.

Well control can be a complex topic, but it is clearly critical to a successful drilling operation. Well-control procedures should be part of well planning, so that the proper actions will be established and crews will be familiar with them when drilling begins. It is essential that rig crews be trained to react quickly and appropriately to an unexpected event that might jeopardize the well.

Completions and Cementing

In drilling terminology, a well’s *completion* refers to the combination of casing, casing accessories, and cement that allows the well to produce. Casing accessories can refer to perforated or slotted liners in the production zone, internal or external packers, or the hardware necessary for multilateral wells that have more than one leg feeding into the production casing.

Cementing

As described in the Drilling Overview, casing has been traditionally cemented in place by pumping a calculated amount of cement into the casing, placing a movable plug on top of the cement, and then

displacing the plug downward with drilling fluid. This forces the cement to flow out the bottom of the casing and up the annulus between the casing and wellbore. In most oil and gas wells, the casing is cemented in place only at the bottom, with a completion fluid between the balance of the casing and the wellbore wall, but geothermal wells must have a complete cement sheath from bottom to surface [55]. This cement has two important functions: to give the casing mechanical support under its sometimes-intense thermal cycling between production and shut-in, and to protect the outside of the casing from corrosion by in situ fluids.

This implies that geothermal cements should have high bond strength to the casing and should be impermeable, but it is also very advantageous for the cement to be lightweight (at least compared to conventional cement, which has a specific gravity of approximately 1.6). Light weight is important because of the oft-encountered lost circulation described above. If the formation’s pore pressure will not even support drilling fluids, then it is impossible to lift a column of normal-weight cement back to the surface when casing is cemented in place. One solution to this problem is foam cement, which has gas injected into it, in the same way as drilling fluid is aerated to make it lighter. Recent experience with difficult wells in California [56] and Hawaii [57] has also shown that reverse circulation foam cementing, where the cement is pumped down the annulus and flows back up drill pipe from the bottom of the casing, has several advantages.

If cement returns have not reached the surface in a conventional cement job (i.e., there is an uncemented annulus around the top of the casing), then some method must be used to cement this remaining volume. The most common method is to pump cement, under pressure, into the top of the annulus, which will fracture the formation down to the top of the existing cement. If the cement has reached as high as the top centralizer, then a “top job,” which usually means that small-diameter lines (tremie lines) are inserted into the annulus and cement is pumped into them to fill the annular volume, can be performed. The risk in this is that liquid (water or drilling fluid) will be trapped between the upper and lower volumes of cement (see below in *Completions*), so all possible precautions should be taken to avoid

this. If the resources are available, the annulus can be dried with steam [58] to assure the absence of liquid.

Conventional oil well cements are not only too heavy for many geothermal wells, but are susceptible to attack by acids and by CO₂, both of which are common in geothermal reservoirs and both of which degrade the impermeability and strength of the cement. Historically, the major modification to Portland cement for geothermal use is the addition to standard Class G cement of retardants and approximately 40%, or more, silica flour [59], but this does not eliminate the problem of CO₂ and acid attack. Brookhaven National Laboratory carried out a major research program on geothermal cement, intended to mitigate or eliminate these effects. The R&D effort comprised: characterization of cements then used in geothermal environments [60, 61], the extension of hydrothermal cements to higher operating temperatures [62], and the development of new materials such as phosphate-bonded cements [63], polymer cements [64], and other new compositions [65].

BNL worked with cost-sharing industry partners (Halliburton, Unocal, and CalEnergy Operating Company) toward the specific goal of a lightweight cement with outstanding resistance to acid and CO₂ at brine temperatures up to 320°C. Reviews of this work before [66] and after [67] 1997 are provided in detailed reports. BNL succeeded in synthesizing, hydrothermally, two new cements: calcium aluminate phosphate (CaP) cement and sodium silicate-activated slag (SSAS) cement. The CaP cements were designed as CO₂-resistant cements for use in mildly acidic (pH ~ 5.0) CO₂-rich downhole environments. The SSAS cements were designed to resist a hot, strong acid containing a low level of CO₂. Both of these were economical cements because they used inexpensive cement-forming by-products from coal combustion and steel-manufacturing processes. In 1997, Unocal and Halliburton completed four geothermal wells in Sumatra with CaP cement, the first field use of this formulation, and in 1999, Halliburton commercialized it under the name “ThermaLock Cement.” SSAS cements have received less attention than CaP, but autoclave experiments in the lab have demonstrated good performance in high-acid environment and, in fact, after undergoing acid damage, the SSAS cement exhibited a self-repairing characteristic. Addition of

fly-ash further improved its acid resistance, so SSAS is promising as low-cost geothermal well cement in high-acid conditions up to 200°C.

Completions

Apart from the requirement for a complete cement sheath around the casing, factors that influence completion design include brine chemistry, multibranch completions, and whether the production interval is stable enough to be open hole or must be completed with a slotted liner.

Brine chemistry can cause two major problems: corrosion and scaling. Brine quality varies greatly, ranging from near-potable in moderate-temperature systems to highly corrosive with high dissolved solids in some high-temperature systems. Many techniques – cement-lined casing, exotic alloys, and corrosion-resistant cement – have been applied to the casing corrosion problem, which is especially severe in the Imperial Valley, California. Shallow and hot, CO₂ bearing zones there drive an external corrosion rate approaching 3 mm of carbon steel per year, so the wells must be abandoned after 10–12 years even after well life is extended by cementing in smaller production strings. Most existing production wells in the Imperial Valley have been completed or retrofitted with titanium casing, which has proved to be cost effective in spite of its very high capital investment (approaching US\$3,000/m.)

Scaling, the buildup of mineral deposits both inside the casing and in the production interval, is a problem in geothermal plants around the world [68], and can lead to frequent workovers. In severe cases, untreated scaling can reduce the flow area of casing by half in a matter of months. Casing scale can sometimes be removed with high-pressure jets [69], but scaling in the wellbore often seals the formation and must be drilled out with an under-reamer (an expandable bit that can drill a hole below casing that is larger than the inside diameter of the casing). It is highly preferable to inhibit or prevent scale formation than to remove it, and there are many chemical techniques for this, but discussion of those is beyond the scope of this article.

When casing is cemented, it is also critical that no water be trapped between the cement and the casing, especially in intervals where one casing is inside another,

because the water can become hot enough as the well goes on production and heats up that thermal expansion can collapse the inner casing. If the trapped-water location has formation outside it, the fracture gradient is usually low enough to allow the pressure to bleed off into a fracture. These failures can be serious enough that the production casing is imploded and ruptured to the extent that it will reduce production and will provide a path from the formation into the cased hole [70].

Finally, it is necessary to decide whether the production interval of the well is competent enough formation so that it can be left as-drilled (open-hole completion) or whether a slotted liner will be necessary to protect against sloughing or caving into the wellbore. Some indications can be gained from the geologic samples acquired during drilling, or from imaging logs, if available, but this decision is often made based on experience gained from other wells in the same reservoir.

Instrumentation (Drilling and Mud Logging)

This description of instrumentation deals only with those measurements applied to the drilling process, and does not address logging for formation evaluation done during or after drilling. Drilling information comprises both surface measurements – those taken on or around the drill rig – and downhole data retrieved by some type of logging tool that is either lowered into the borehole or forms a part of the BHA.

Surface Measurements

A summary list of desirable measurements for the drill rig was given in the section on rig selection criteria (drilling fluid inflow and outflow rates, drilling fluid inflow and outflow temperatures, standpipe pressure, rotary speed, weight on bit, and torque) but many others exist. The drill rig will have at least a minimum set of instruments that are required for its normal functions, but additional instrumentation and data can be provided by the drilling contractor, the mud-logging company (MLC), or an independent service company. It is most commonly done by the MLC, in conjunction with their primary job of recording the geology of the well, based on the cuttings brought back to surface by the drilling fluid. The MLC also keeps a record of many of the drill rig's operating conditions – a representative MLC, for example, lists all the

following measurements as available, so it is the well planner's responsibility to decide which are necessary.

- Depth
- Block height
- Rate of penetration
- Bit depth tracking while tripping
- On bottom/off bottom
- Hook load
- Weight on bit
- Rotary RPM and torque
- Top drive RPM and torque
- Standpipe pressure
- Casing pressure
- Pump stroke rates
- Pump stroke counters
- Totalized pit volumes
- Individual pit volumes
- Trip tank volumes
- Mud gain/loss
- Mud flow rates
- Mud temperature in and out
- Mud weight in and out
- Mud resistivity in and out
- CO₂ and H₂S

Understanding how to use the measurements is clearly important, and should be part of the driller's training. Some comparisons, such as mud flow rates in and out of the wellbore, have been described previously as diagnostics for lost circulation and/or well-control issues. Others, such as a sudden drop in standpipe pressure as an indication of a washout in the drill string, should be part of training. Many of the measurements made by the MLC can be combined electronically in such a way that an alarm will sound if undesirable conditions appear (e.g., the difference in flow rates becomes large). Virtually, all modern MLCs present and record data in digital format, so that it is easily stored, retrieved, and displayed at multiple locations (including a web site, if desired.)

It is also possible to use longer-term data – torque and weight on bit related to rate of penetration, pump efficiency compared to mud flow rate, temperature change as a function of depth – to establish statistical trends that are a measure of drilling performance or downhole conditions. It is also possible, in principle, to combine surface measurements in a way that provides diagnostics for

various drilling conditions and then employs an expert system approach to recommend subsequent action. This has been investigated in the laboratory [71] and some versions of it have been commercialized.

Downhole Measurements

Surface measurements are often ambiguous because there is more than one downhole condition that can produce the same readings at the surface, so downhole measurements are highly valuable in resolving this discrepancy. Downhole measurements can be made in several different ways:

- A sensor package can be lowered into the hole on an electrically conducting cable (wireline), sending back signals in real time as it traverses the wellbore. This method usually requires a specialized wireline truck operated by a logging service company (i.e., this method is relatively expensive and there is some lead time involved unless the truck and crew are on standby at the drill site). Real-time information is advantageous when a very dynamic situation such as drilling is in progress, especially if there is reason to believe that some downhole condition (e.g., pressure, lost circulation, bit dysfunctions) may be harmful, hazardous, or expensive.
- A logging tool with onboard memory can be lowered into the hole on an ordinary cable (slickline), taking readings as it traverses the wellbore, and then brought back to surface where data is downloaded. If real-time data is not required, this method tends to be cheaper and more convenient, because the memory tool can be operated by the rig crew on the rig's hoisting equipment.
- A memory tool can also be part of the BHA, retrieved either when tripping the drill string or by slickline. This method is particularly useful when a slimhole is being drilled with a coring rig, because the memory tool can be part of the core tube and data can be retrieved with every core run.
- An instrumentation package that is part of the BHA can send signals back to the surface through pressure pulses in the drilling fluid. This "mud-pulse telemetry" is most often used for directional drilling, where it provides survey information for steering the hole's trajectory, but it can also send back information on downhole conditions such as

pressure and temperature, or on drilling parameters such as shock and vibration. This method provides real-time data from the bottom of the hole, but has a number of disadvantages, in that it is very expensive, is susceptible to high temperature, cannot operate in aerated mud or air, and has a very low data rate (less than 10 baud).

- Signals can also be sent back to surface from a near-bit instrument package through stress waves in the steel drill pipe. This "acoustic telemetry" is reasonably rugged, has a higher data rate than mud pulse (above 20 baud), can operate in any drilling fluid, and has been commercialized by a company in Canada [72].

All of this technology is very mature for the oil and gas industry, but high temperature is a barrier for much geothermal work. Although other parts of a downhole instrumentation package (e.g., seals, the wireline cable head, and sensors) become more difficult in high temperatures, electronic components are the principal challenge. Commercially available electronic components are generally rated at only about 85°C (This is a "guaranteed" operational temperature, although selected components will operate at higher temperatures), unsuitable for use in geothermal environments, so there are three choices: (1) develop electronic components that can withstand higher temperatures, (2) shield conventional components from the high-temperature environment, or (3) use a combination of (1) and (2).

Electronic components can be protected from high temperature by enclosing them in a thermal flask, or Dewar. A Dewar functions like a Thermos bottle, with an evacuated volume between concentric shells providing insulation for the components inside. Like a Thermos bottle, a Dewar in a hot well will eventually (the length of time that the Dewar will protect the electronics is a function of the wellbore temperature, power dissipation requirements of the electronics package, conductivity of the Dewar, and the heat sink inside the package. For typical geothermal applications, the operating envelope is 6–16 h) allow the components inside to heat up to a point at which they may fail. Dewars provide only temporary protection and are expensive and fragile, but even when using high-temperature electronics, they will give the logging tool additional life. Almost all logging tools, both wireline and memory, used in geothermal environments are protected by Dewars.

Electronic components that can operate, unprotected by thermal flasks, at geothermal temperatures, are the ultimate goal. Two technologies – silicon-on-insulator (SOI) and silicon carbide (SiC) – approach that goal. SOI semiconductors can operate virtually indefinitely [73] at 300°C; SiC semiconductors above 450°C – well above existing electronic packaging technology. Some SOI electronic components have been commercially available for several years [74], and a basic suite of SOI-based logging tools is commercially available now.

Of the many measurements that can be made by logging downhole, by far the most useful for drilling purposes is temperature. Apart from the clear necessity to know whether the hole is approaching a geothermal resource, temperature logs can clarify a number of other drilling situations.

- Logs can provide warning if any temperature-limited downhole equipment (including drilling fluid) is approaching its limit.
- If a lost-circulation zone appears during drilling, logs can often define its location (it is not always at the bottom of the hole).
- Logs can guide the amount of retarder to add to cement before cementing casing.
- Because cement has an exothermic reaction as it cures, logs can locate the top of the cement column if returns do not reach the surface.
- For an injection test in a potential production zone, logs can identify fracture locations.
- Logs can usually identify favorable (impermeable) zones for setting packers, if that is part of the test program.

This is only a sample of the applications that logs can have and, for the cases cited above, real-time data is not critical, so memory tools would be quite adequate. These examples indicate the versatility of temperature logs, so having this capability as a standard part of the drilling program is highly useful.

Future Directions

The future direction of geothermal drilling is, in many ways, undefined. This uncertainty stems from the multiple development scenarios that can be envisioned for the industry. It is widely believed that enhanced geothermal systems (the EGS concept is described in detail elsewhere

in this encyclopedia but, in general, it means injecting fluid into one well or set of wells, forcing that fluid to gain heat by circulating through fractures in the hot reservoir, and then returning it to the surface through another well or wells. Unlike conventional hydrothermal resources, EGS wells do not produce in situ fluids) (EGS) will provide the bulk of new geothermal capacity, worldwide, but many aspects of EGS development are unresolved. Resource location, reservoir creation, and reservoir management will all be different when applied to EGS than is the case in conventional hydrothermal practice, and these differences could well drive drilling research and development in a new direction compared to past R&D for hydrothermal resources.

Costs and risks may also follow a different pattern with EGS. It is well known that geothermal wells cost more than oil and gas wells of comparable depth, and that drilling costs increase more than linearly with depth. Costs for deep geothermal wells, however, do not increase as rapidly with depth as costs for deep oil and gas wells [36]. There are, in general, four ways to reduce well cost: eliminate “trouble” costs, improve the efficiency of standard operations, introduce new and more efficient operations, or change the well design [75]. Because the “average” EGS well is expected to be considerably deeper than the “average” hydrothermal well, however, the focus of cost reduction may shift among these priorities.

Regardless of the directions that drilling research and EGS development may follow, it is still likely that the geothermal drilling market will remain so small, relative to the oil and gas market, that most innovation in geothermal drilling will derive from technology used in the oil patch. Given that assumption, it is useful to look at several drilling methods and technologies that have gained wide acceptance in the oil and gas industry but have been applied sparingly, if at all, in geothermal wells. The following section describes these technologies, summarizes their advantages (with focus on the geothermal context), and discusses the barriers to their use in geothermal drilling.

Drilling with Casing (DWC)

The casing can be used as the drill string, rotating to turn a bit and advancing with the hole as it gets deeper, so that it is already in place when the hole reaches

desired depth. There are two basic ways that the bit can be attached to the casing: it can be semipermanently mounted, so that it can be dropped off the end of the casing at final depth, or can be drilled through for passage of a subsequent casing string; or it can be mounted on a drilling assembly that is retrieved either by wireline or drill pipe when the bit needs to be changed, or when the hole is at design depth. If a retrievable bit is used, then it must be small enough to pass through the casing's inside diameter; therefore, it must use an under-reamer to cut the diameter that is large enough for the outside of the casing couplings and the annulus for the drilling fluid return flow.

The casing must always be rotated by a top drive unit, and can be connected to the top drive by either screwing into the top coupling of the casing or by a fixture that stabs into the top joint of casing and locks to its inside diameter. The top drive circulates drilling fluid through the casing's inside and back up the outside, just as it would with drill pipe. As in conventional use, the top drive also has the ability to circulate continuously, which can be important in geothermal drilling with heat-sensitive downhole tools.

There are several advantages to this technology, as described in the cited reference [76].

- Eliminate costs, time, and problems related to tripping drill pipe – Time to trip drill pipe and handle the BHA is a significant fraction of total time (and cost) on some wells [77], but it is also the case that many problems of well control and hole stability are associated with trips.
- Reduce lost-circulation problems – Drilling with casing (DWC) systems can continue drilling when lost circulation is encountered. The rock cuttings tend to be washed into the fractures or permeable zones, acting effectively as lost-circulation material. The relatively narrow annulus also means that fluid flow rates can be lower than would be used with conventional drilling in the same size hole.
- Gain casing setting depth – The ability to drill through lost-circulation zones, or other weak formations, means that sometimes the casing can reach a greater depth than would be the case with conventional drilling. It is possible, for some well designs and lithologies, that the casing program could be redesigned to eliminate one string of

casing. As shown in Section [Well Cost](#), this is a major saving.

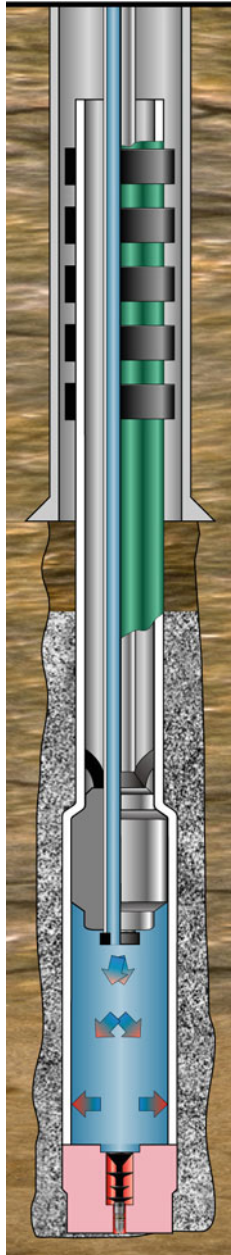
- Improve safety – Handling drill pipe has one of the highest accident incidences in drilling; eliminating this activity means that the crew is exposed to less risk.

Although this technology has been used on hundreds of oil and gas wells, it has seen very limited use for geothermal. There is an issue with retrievable drilling assemblies, because they contain some elastomer components, but the larger factor is hard rock. The cutting structure for most DWC bits uses PDC cutters and, as discussed previously, these cutters have not been notably successful in geothermal formations. Some field experience with hard rock in oil and gas drilling, however, indicates that reasonable performance with roller-cone bits and PDC under-reamers is available [78]. Although a number of questions remain to be answered, this technology appears to have enough potential to warrant further investigation devoted specifically to geothermal drilling.

Expandable Tubulars

As described earlier, casing is installed in successively smaller diameters (see [Fig. 3](#)) as the hole gets deeper, so that maintaining the correct diameter in the production zone means having much larger holes and casing at the top of the well. It should also be noted that there is a sizable difference in diameter (10–20 cm) between successive casing strings, so that in the example figure, a 21.6 production interval requires drilling a 102 cm hole at the top. This difference in diameter is required to allow clearance for the couplings on the outside of the inner casing string, to compensate for the fact that the previous casing may not be in a straight hole, and to give sufficient annular area that cement can easily flow through it.

The larger casing sizes and cementing jobs at the top are expensive, however, and drilling larger diameter holes often is slower than drilling a smaller diameter would be. A relatively new technology (first field tested in 1998) makes it possible to run a string of casing with normal clearances and then expand the diameter of the inner string so that the clearance between the two strings is negligible. This diameter increase is implemented by an “expansion cone” in the bottom of the inner casing string (see [Fig. 12](#)). Once the hole is drilled, the liner, with the cone assembly in the bottom



Geothermal Resources, Drilling for. Figure 12
Expandable liner (Diagram courtesy of Enventure Global Technology)

joint, is made up until the desired length is complete. Drill pipe is then screwed into the cone launcher assembly and the liner is run into the hole on the drill pipe. Cement is pumped in the normal way (except less volume than would normally be used) and the cone is forced up the liner by a combination of hydraulic

pressure beneath it (delivered through the drill pipe) and pulling with the drill pipe. As the liner expands, it forces the cement upward until the liner annulus is completely cemented.

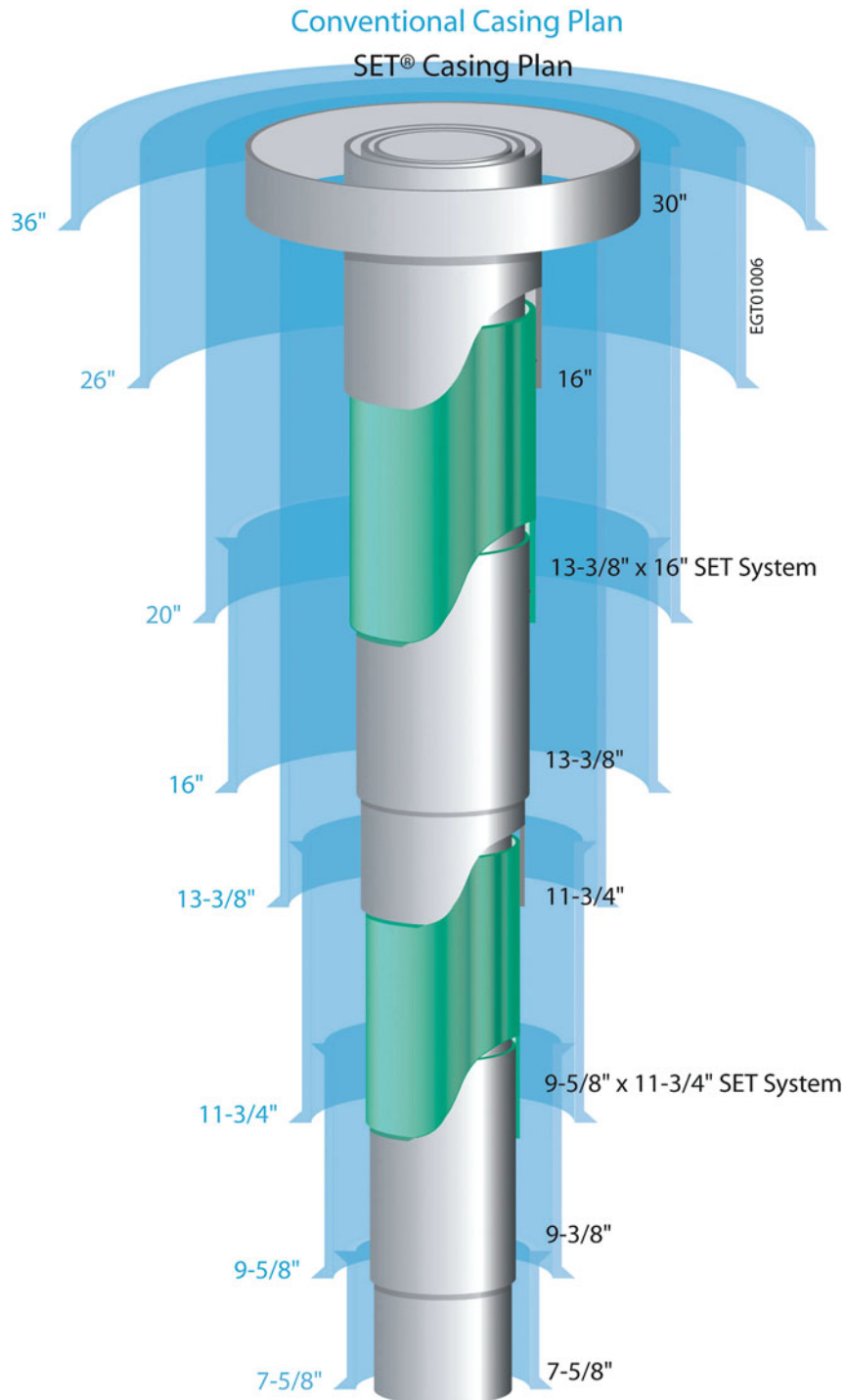
As shown in Fig. 13, using this system (called SET – solid expandable tubulars – by one manufacturer) means that much less clearance between successive casing strings is necessary and, therefore, the upper casings can be smaller for a given production zone diameter than with conventional casing [79].

When considering this system for geothermal drilling, there are at least two potential vulnerabilities: the expandable tubulars depend on elastomer seals in some applications, so the temperature rating of the seals is critical; and the inside of the casing has a proprietary coating to ease the cone's passage through the pipe. It is not clear how this coating would be affected by high temperature, but the system has been used in the field [80] at temperatures above 160°C, and tests are under way to qualify seals for use above 250°C in steam-enhanced oil recovery (SAGD).

Another possible use for expandables is to repair or mitigate lost circulation. A section of casing can be expanded into open hole, rather than into a previous casing string and, if the open hole section has been slightly under-reamed, there will be little, if any, loss of diameter because of the patch. Since no cement is used in this treatment, the casing depends entirely on its external elastomers for zonal isolation, making this component especially important for this application. If it can be shown that all components of expandable tubulars can withstand high temperature, then it appears that there are at least two valuable applications for expandable tubulars in geothermal drilling.

Better Downhole Feedback

Downhole measurements are mentioned in the Instrumentation section, but those were generally measurements related to the state of the wellbore (deviation and trajectory), not real-time data for drilling performance. A few exceptions exist; there is experience with using mud-pulse telemetry to send vibration data uphole [81], but the low mud-pulse data rate limits the applications of these systems. Low data-transmission rates also mean that the downhole sensor package must include a great deal of data processing and, as described



Geothermal Resources, Drilling for. Figure 13

Comparison of casing diameters between SET technology and conventional casing (Diagram courtesy of Enventure Global Technology)

previously, keeping the electronics functional at geothermal temperatures is a challenge.

Better real-time data collection, transmission, and interpretation is a high priority in drilling, corroborated by an industry forum [82] that identified this as the most important technology need for reducing flat time (defined as the time the rig is over the hole, with the hole not advancing). The principal barrier to much of this activity has been the lack of a transmission method with adequate bandwidth. In the last decade, however, drill pipe with built-in instrumentation cable has been developed [83]. This pipe has been used at high bandwidth in the field [84], although not at geothermal conditions, and it offers a very promising opportunity for expanded use of real-time downhole data.

The list of measurements that can be made and transmitted is, of course, very large but they fall into three general categories:

- Improve drilling performance – One of the most common causes for poor bit performance, especially with PDC bits, is excessive shock and vibration. Sandia National Laboratories developed a high-data-rate downhole sensor package to improve PDC performance, and field tests [85] with and without the package demonstrated that it significantly lengthened the life of a PDC bit in hard rock. The primary benefit of the real-time data system was to allow destructive downhole conditions to be immediately recognized and mitigated, but it also showed that surface readings for some parameters (e.g., weight on bit) were much different from the values actually measured near the bit. A later version of this system, modified for high temperature, was run in a geothermal well [86], but it has not been commercialized. Other bit dynamics packages are commercially available, although generally not at high bandwidth.
- Avoid trouble – Aside from the problem of bit failure, many other kinds of trouble (well control, lost circulation, unexpectedly high temperature) can either be avoided or recognized much earlier, allowing more effective treatment, with real-time data.
- Eliminate logging time – With properly configured downhole packages, the well can be logged as it is

drilled, eliminating the time (sometimes days) at the end of an interval, or the end of the well, normally required to log it. In some cases, logging could be done with a memory tool rather than with real-time instrumentation, but there is a risk that failure in the logging tool would go undetected until drilling was over.

All of these uses imply high-bandwidth transmission systems and, for geothermal drilling, high-temperature downhole electronics (as well as high-temperature batteries). All of these technologies exist in some form, but they have not yet been put together for geothermal drilling.

For More Information

For more information about the topics in this entry, apart from the extensive references cited, other resources are available. Both the Society of Petroleum Engineers (www.onepetro.org) and the Geothermal Resources Council (www.geothermal.org) provide searchable databases of their own publications that include detailed descriptions of geothermal drilling technology. All of the cited references from *Geothermal Resources Council Transactions* are available through the GRC web site (free to members, nominal charge to nonmembers). Stanford University hosts an annual Geothermal Workshop, and papers from those meetings, as well as from World Geothermal Congress, can be located through <http://pangea.stanford.edu/ERE/db/IGASTandard/search.htm>. The Office of Science and Technology Information maintains the Department of Energy's Geothermal Technologies Legacy Collection (<http://www.osti.gov/geothermal/>) and many of the papers cited in this entry are available through that resource. The US Bureau of Land Management provides a summary document describing regulatory requirements for exploration, drilling, production, and abandonment on Federal geothermal leases [87] and The Standards Association of New Zealand has published a 93-page manual that combines regulatory requirements with suggestions on operational practices for drilling, maintenance, repair, and abandonment [88]. Finally, the oil-field service company Schlumberger maintains an online glossary with definitions of many common drilling terms at <http://www.glossary.oilfield.slb.com/>.

Bibliography

1. International Geothermal Association, Installed Generating Capacity. <http://www.geothermal-energy.org/geoworld/geoworld.php?sub=elgen>
2. US Department of Energy, Energy Information Administration. http://tonto.eia.doe.gov/dnav/pet/pet_crd_wellend_s1_a.htm
3. Burgoyne AT Jr et al (1986) *Applied drilling engineering*. Society of Petroleum Engineers, Richardson
4. Mitchell RF (ed) (2007) *Petroleum engineering handbook, volume II: drilling engineering*. Society of Petroleum Engineers, Richardson. ISBN:978-1-55563-114-7
5. Cacini P, Mesini E (1994) Rock-bit wear in ultra-hot holes SPE 28055. SPE/ISRM rock mechanics in petroleum engineering conference, Delft, Netherlands
6. Holligan D et al (1989) Performance of beta titanium in a Salton Sea geothermal production well. SPE 18696, SPE/IADC drilling conference, New Orleans, LA
7. Mansure AJ (2002) Polyurethane grouting geothermal lost circulation zones. SPE 74556 IADC/SPE drilling conference, Dallas TX
8. Renner JL et al (2007) *Geothermal engineering*. In: Warner HR Jr, (ed) *Petroleum engineering handbook, vol. VI: emerging and peripheral technologies*. ISBN 978-1-55563-122-2
9. Saito S, Sakuma S (2000) Frontier geothermal drilling operations successful at 500°C BHST, SPE 65104, SPE drilling and completion, September 2000
10. Pierce KG, Livesay BJ (1994) A study of geothermal drilling and the production of electricity from geothermal energy, SAND92-1728, Sandia National Laboratories
11. Pierce KG, Bomber TM, Livesay BJ (1997) Well cost estimates in various geothermal regions. *Geothermal Resources Council Transactions*, vol. 21
12. Mansure AJ et al (2001) Polyurethane grouting of Rye Patch lost circulation zone. *Geothermal Resources Council Transactions*, vol. 25
13. Holligan D, Cron CJ, Love WW, Buster JL (1989) Performance of beta titanium in a Salton Sea field geothermal production well source, SPE18696, SPE/IADC Drilling Conference, New Orleans, LA
14. Finger JT, Jacobson RD, Hickox CE (1997) Newberry exploratory slimhole: drilling and testing SAND97-2790, Sandia National Laboratories
15. Finger JT, Jacobson RD, Hickox CE (1996) Vale exploratory slimhole: drilling and testing SAND96-1396, Sandia National Laboratories
16. Finger JT et al (1999) Slimhole handbook: procedures and recommendations for slimhole drilling and testing in geothermal exploration SAND99-1976, Sandia National Laboratories
17. <http://office.microsoft.com/en-us/project/default.aspx>
18. Combs J, Garg SK, Livesay BJ (2000) Maximum discharge of geothermal fluids from slim holes by optimizing casing designs. *Geothermal Resources Council Transactions*, vol. 24
19. Mitchell RF (ed) (2007) *Petroleum engineering handbook, volume ii: drilling engineering*. Society of Petroleum Engineers, Richardson. ISBN:978-1-55563-114-7
20. Cavanaugh, J.M., Adams, D.M., (1988). Top-drive drilling system evaluation. *SPE Drill Eng* 3(1) 43–49
21. Saito S et al (2003) Advantages of using top-drive system for high temperature geothermal well drilling. *Geothermal Resources Council Transactions*, vol. 27, pp 183–187
22. US Patent 930,758 "Drill" issued 10 Aug 1909
23. Finger JT, Glowka DA (1989) PDC bit research at Sandia National Laboratories, Sandia Report SAND89-0079, Sandia National Laboratories
24. Wise JL et al (2003) Hard-rock drilling performance of a conventional PDC drag bit operated with, and without, benefit of real-time downhole diagnostics. *Geothermal Resources Council Transactions*, vol. 27
25. Hareland G et al (2009) Cutting efficiency of a single PDC cutter on hard rock. *J Can Pet Technol* 48(6):60–65
26. Rickard WM, Johnson B, Mansure AJ, Jacobson RD (2001) Application of dual tube flooded reverse circulation drilling to rye patch lost circulation zone. *Geothermal Resources Council Transactions*, vol. 24
27. Zilch HE, Otto MJ, Pye DS (1991) The evolution of geothermal drilling fluid in the Imperial Valley SPE21786, Presentation at the SPE Western Regional Meeting, Long Beach, CA
28. Carter TS (ed) (1997) *Drilling fluids*. Society of Petroleum Engineers, Richardson. ISBN:978-1-55563-069-0
29. Bourgoyne Jr AT, Millheim KK, Chenevert ME & Young Jr FS (1991) *Applied Drilling Engineering*, Society of Petroleum Engineers, ISBN:978-1-55563-001-0
30. American Petroleum Institute (2003) RP 13B-1/ISO 10414-1, Recommended practice for field testing water-based drilling fluids (includes Errata, July 2004) Product Number: GX13B13
31. Tuttle JD (2005) Drilling fluids for the geothermal industry – recent innovations. *Geothermal Resources Council Transactions*, vol. 29
32. Jaimes-Maldonado J, Cornejo-Castro S (2006) Case study: underbalanced or mud drilling fluids at Tres Virgenes geothermal field. *Geothermal Resources Council Transactions*, vol. 30
33. Finger JT, Jacobson RD, Hickox CE, Combs J, Polk G, Goranson C (1999) Slimhole handbook: procedures and recommendations for slimhole drilling and testing in geothermal exploration, Sandia Report SAND99-1976, Sandia National Laboratories
34. Carson CC, Lin YT (1982) the impact of common problems in geothermal drilling and completion. *Geothermal Resources Council Transactions*, vol. 6
35. Rickard WM et al (2001) Application of dual tube flooded reverse circulation drilling to Rye Patch lost circulation zone. *Geothermal Resources Council Transactions*, vol. 25
36. Mansure AJ, Bauer SJ (2005) Advances in geothermal drilling technology: reducing cost while improving longevity of the well. *Geothermal Resources Council Transactions*, vol. 29

37. Petty S et al (2005) Lessons Learned in Drilling DB-1 and DB-2, Blue Mountain NV, Proceedings, thirtieth workshop on geothermal reservoir engineering. Stanford University, CA
38. Drotning WD, Ortega A, Harvey PE (1982) Thermal conductivity of aqueous foam, Sandia Report SAND82-0742, Sandia National Laboratories
39. Rand PB, Montoya O (1983) Evaluation of aqueous foam surfactants for geothermal drilling fluids, Sandia Report SAND83-0584, Sandia National Laboratories
40. Gislason T, Richter B (2008) The aerated drilling experience of icelandic geothermal wells. Geothermal Resources Council, Transactions vol. 32
41. Loeppke G (1986) Evaluating candidate lost circulation materials for geothermal drilling. Geothermal Resources Council Transactions, vol.10
42. Sugama T et al (1986) Bentonite-based ammonium polyphosphate cementitious lost-circulation control materials. J Mater Sci 21
43. Staller GE (1999) Design, development and testing of a drillable straddle packer for lost circulation control in geothermal drilling, Sandia Report SAND99-0819, Sandia National Laboratories
44. Glowka DA et al (1989) Laboratory and field evaluation of polyurethane foam for lost circulation control. Geothermal Resources Council Transactions, vol. 13
45. Mansure AJ, Westmoreland JJ (1999) Chemical grouting lost-circulation zones with polyurethane foam. Geothermal Resources Council Transactions, vol. 23
46. Mansure AJ et al (2001) Polyurethane grouting of rye patch lost circulation zone. Geothermal Resources Council Transactions, vol. 25
47. Mansure AJ et al (2004) Polymer grouts for plugging lost circulation in geothermal wells, Sandia Report SAND2004-5853, Sandia National Laboratories
48. Schafer DM et al (1992) Development and use of a return line flowmeter for lost circulation diagnosis in geothermal drilling. Geothermal Resources Council Transactions, vol. 16
49. Whitlow GL et al (1996) Development and use of rolling float meters and Doppler flow meters to monitor inflow and outflow while drilling geothermal wells. Geothermal Resources Council Transactions, vol. 20
50. Manual M07, California Department of Conservation (2006) Blowout Prevention in California. http://www.conservation.ca.gov/dog/geothermal/pubs_stats/Pages/instruction_manuals.aspx
51. Karner SL (2005) Creating permeable fracture networks for EGS: engineered systems versus nature. Geothermal Resources Council, Transactions, vol. 29
52. New Scientist (1991) Blowout blights future of Hawaii's geothermal power, 20 July 1991, issue 1778
53. Herras EB et al (2004) A geoscientific approach in the design and success of the first relief well at the Leyte geothermal production field. Geothermal Resources Council Transactions, vol. 28
54. Pye DS, Hamblin GM (1991) Drilling geothermal wells at the geysers field. Geothermal Resources Council, Monograph on the Geysers Geothermal Field, Special Report No. 17
55. Nelson EB et al (1981) Evaluation and development of cement systems for geothermal wells SPE10217, Society of Petroleum Engineers
56. Bour DL, Hernandez R (2003) CO₂ Resistance, improved mechanical durability, and successful placement in a problematic lost circulation interval achieved: reverse circulation of foamed calcium aluminate cement in a geothermal well. Geothermal Resources Council Transactions, vol. 27
57. Spielman P et al (2006) Reverse circulation of foamed cement in geothermal wells. Geothermal Resources Council Transactions, vol. 30
58. Saito S (1994) A new advanced method for top-job casing cementing. Geothermal Resources Council Transactions, vol. 18
59. Koons BE et al (1993) New design guidelines for geothermal cement slurries. Geothermal Resources Council Transactions, vol. 17
60. Nelson EB Development of geothermal well completion systems, final report, Dowell Division, Dow Chemical, USA, DOE Contract DE-ACO2-77ET28324
61. Kalyoncu RS, Snyder MJ (1981) High-temperature cementing materials for completion of geothermal wells, BNL-33127, Brookhaven National Laboratory
62. Curtice DK, Mallow WA (1979) Hydrothermal cements for use in the completion of geothermal wells, Southwest Research Institute, BNL 51183
63. Rockett TJ (1979) Phosphate-bonded glass cements for geothermal wells, University of Rhode Island, BNL 51153
64. Zeldin AN, Kukacka LE (1980) Polymer cement geothermal well-completion materials, final report, Brookhaven National Laboratory, BNL 51287
65. Roy DM et al (1980) New high temperature cementing-materials for geothermal wells: stability and properties, The Pennsylvania State University, BNL 51249
66. Kukacka L (1997) Geothermal materials development at Brookhaven National Laboratory, BNL-64482, Brookhaven National Laboratory
67. Sugama T (2006) Advanced cements for geothermal wells, BNL 77901-2007-IR, Brookhaven National Laboratory
68. Ocampo-Díaz J, Rojas-Briebesca M (2004) Production problems review of Las Tres Virgenes geothermal field, Mexico. Geothermal Resources Council Transactions, vol. 28
69. Hurtado R, Mercado S (1990) Scale control studies at the Cerro Prieto geothermal plant. Geothermal Resources Council Transactions, vol. 14
70. Southon JNA (2005) Geothermal well design, construction and failures. Proceedings World Geothermal Congress, Antalya, Turkey

71. Harmse JE et al (1997) Automatic detection and diagnosis of problems in drilling geothermal wells. *Geothermal Resources Council Transactions*, vol. 21
72. Drumheller DS, Kuszmaul SS (2003) Acoustic telemetry, Sandia Report SAND2003-2614, Sandia National Laboratories
73. Normann RA, Henfling JA (2004) Aerospace R & D benefits future geothermal reservoir monitoring. *Geothermal Resources Council Transactions*, vol. 28
74. Henfling JA, Normann RA (2002) Advancement in HT electronics for geothermal drilling and logging tools. *Geothermal Resources Council Transactions*, vol. 26
75. Mansure AJ et al (2005) Geothermal well cost analyses 2005. *Geothermal Resources Council Transactions*, vol. 29
76. Warren TM (2009) Casing while drilling, in advanced drilling and well technology, Society of Petroleum Engineers, ISBN 978-1-55563-145-1
77. Polsky Y (2008) Enhanced geothermal systems (EGS) well construction technology evaluation report, SAND 2008-7866, Sandia National Laboratories
78. Tessari RM, Warren TM (2003) Casing drilling reduces lost circulation problems. *Geothermal Resources Council Transactions*, vol. 27
79. Tubbs D, Wallace J (2006) Slimming the wellbore design enhances drilling economics in field development, SPE 102929, SPE Annual Conference and Exhibition, San Antonio, TX
80. Nylund J et al (2009) Integrating solid expandables, swellables, and hydra jet perforating for optimized multi-zone fractured wellbores, SPE 125345, Tight Gas Completions Conference, San Antonio, TX
81. Shuttleworth NE et al (1998) Revised drilling practices, VSS-MWD tool successfully addresses catastrophic bit/drillstring vibrations, SPE 39314, SPE/IADC Drilling Conference, Dallas, TX
82. Drilling Engineering Association and Energy Research Clearing House (1999) Flat time reduction opportunities: an industry forum, Houston Advanced Research Center, 21 September 1999, Houston, TX
83. Jellison MJ et al (2003) Telemetry drill pipe: enabling technology for the downhole Internet, SPE 79885, IADC/SPE Drilling Conference, Amsterdam, Netherlands
84. Allen S et al (2009) Step-change improvements with wired-pipe telemetry, SPE 119570, SPE/IADC Drilling Conference, Amsterdam, Netherlands
85. Finger JT et al (2003) Development of a system for diagnostic-while-drilling (DWD), SPE 79884, IADC/SPE drilling conference, Amsterdam, Netherlands
86. Blankenship DA et al (2005) High-temperature diagnostics-while-drilling system. *Geothermal Resources Council Transactions*, vol. 28
87. 43 CFR Part 3200, Geothermal resources leasing and operations; final rule, Federal Register, vol. 63, No. 189, 30 September 1998
88. Standards Association of New Zealand (1991) New Zealand Standards NZS 2403:1991, Code of practice for deep geothermal wells, 93 pp

Geothermal Resources, Environmental Aspects of

TREVOR M. HUNT

GNS Science, Wairakei Geothermal Research Centre,
Taupo, New Zealand

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Possible Environmental Impacts and Their Causes
 Methods of Avoiding or Minimizing Impacts
 Future Directions
 Bibliography

Glossary

Aquiclude A geological formation which will not transmit water and is a barrier to vertical movement of geothermal fluid.

Aquifer A geological formation (or formations) which contains water or geothermal fluid and will allow fluid movement.

Downflow Flow of water down a path of high-permeability such as a fracture or a drillhole.

Enhanced geothermal system (EGS) Also called “hot-dry rock”. A form of geothermal development in which heat is extracted from rocks that are hot but have low permeability and often low porosity. The rock is artificially fractured by pumping water into it to form a subsurface heat exchanger.

Epicenter The point on the Earth’s surface directly above the hypocenter or focus of an earthquake.

Groundwater Water, generally cold and of meteoric origin, which resides in near-surface aquifers and is often used for domestic and industrial purposes.

High-temperature system A geothermal system, or part thereof, containing fluid having a temperature greater than 150°C. c.f. *Low-temperature system* in which the temperature is less than 150°C. It can also be argued that the temperature limit be 180°C, since this is the temperature required for a self-discharging well in a liquid-dominated field.

Hydrothermal eruption An eruption resulting from a localized increase in steam pressure in near-surface

aquifers exceeding the overlying lithostatic pressure, and the overburden is then ejected, generally forming a crater 5–500 m in diameter and up to 500 m in depth (although most are less than 10 m deep).

Hypocenter The point at which an earthquake occurs (i.e., the place of rupture), in three-dimensional coordinates (x, y, z) c.f. epicenter (x, y).

Injection (reinjection) The process of returning waste water from a geothermal power station back into the ground. This is generally around the edges of the field and may not be into the production aquifer from which fluid is drawn off to the power station.

Liquid-dominated system A geothermal system, or part thereof, in which the pressure is hydrostatically controlled. c.f. *Steam (vapor)-dominated system* in which the pressure is steam static.

Non-condensable gas A gas present in geothermal fluid which does not become dissolved in the waste water after the fluid has been condensed.

Reservoir The region of a geothermal system from which geothermal fluid is withdrawn, or is capable of being withdrawn.

Permeability A measure of the ability of a geological formation to transmit a fluid.

Waste water Geothermal water from which energy has been extracted and is no longer required. This may be steam that has passed through turbines or a binary plant and been condensed, or separated water.

Definition of the Subject

The Encyclopedia of Environmental Science considers the environment to be “the sum of all external conditions and influences affecting the life and development of organisms” [1] and the Oxford Dictionary defines it as “the set of circumstances or conditions . . . in which a person or community lives, works, develops, etc., or a thing exists or operates; the external conditions affecting the life of a plant or animal” [2].

The term “environment” is therefore generally used in a broad sense to encompass not only physical conditions, but also the cultural and spiritual conditions of people living nearby.

All sources of energy that are used involve some impact on the environment, either in the process of

energy extraction, use or in manufacturing the equipment involved. Key features of the geothermal environment are:

Natural Thermal Features

In many geothermal fields there are beautiful natural thermal features that vary in color and form: geysers, fumaroles, hot springs and pools, silica sinter terraces, mud pools, algal mats, thermophilic plants, and areas of heated ground. They are environmentally important because they are rare on a worldwide basis, and often fragile.

Thermal features are often associated with myths and legends in native peoples culture [3]. For example, the native Maori people of New Zealand have a legend that the thermal areas of NZ were formed when fire gods, summoned from far away and traveling underground, surfaced looking for the person who called them. Many societies that use geothermal energy incorporate it into their ceremonies, for example, in Beppu (Japan) they hold a Hot Spring Festival every year.

Cultural Uses

Bathing in geothermal waters is often claimed to have special medicinal properties, and in New Zealand, geothermal waters are used in the government hospital at Rotorua for the treatment of arthritis and skin diseases. Boiling hot pools are used for cooking: food is placed in a woven basket and lowered into the hot pool – this is still done in Japan and New Zealand, but mainly for tourists. In primitive native societies, red and yellow ochre, formed from hydrothermal alteration of rocks, was used to paint the face and body.

Reasons for Preservation

The most compelling reasons why attempts should be made to preserve the environment are:

Self-respect Most human cultures value their surroundings, even to the extent of significantly modifying them to enhance its beauty or desirability. It is generally recognized that the destruction of beautiful natural thermal features such as geysers, hot springs, and silica terraces is unacceptable. The famous American

philosopher Thoreaux said: “What is the use of a house if you have not got a tolerable planet to put it on?” [4].

Self-preservation Few advanced living organisms will significantly alter or destroy their surroundings because this is likely to threaten their continued existence as a species.

Maintaining Heritage The natural environment is a heritage, inherited from preceding generations, and it is the responsibility of the present generation to pass it undamaged to future generations.

Economic Effects Changing the environment can have negative economic effects. In the case of geothermal development, the destruction, loss, or modification of beautiful natural thermal features can badly affect tourism, which is often a major source of revenue and employment.

To Meet National and International Obligations In most countries, industrial development (including geothermal) is contingent on the developer obtaining a permit (from a regulatory authority) that involves assessing the impact the development may have on the environment and these are difficult to obtain if significant environmental effects are predicted. Preservation of the environment is also of international concern: 21 of 27 Principles proclaimed by the 1992 United Nations Conference on Environment and Development (Earth Summit) refer specifically to the environment.

Introduction

Use of geothermal energy may have some environmental impacts, most of which are associated with the exploitation of high-temperature liquid-dominated geothermal systems for electric power generation. The majority of these impacts, however, can be avoided or minimized with appropriate techniques.

Possible Environmental Impacts and Their Causes

Impact of Access and Field Development Destruction of forests and vegetation resulting from construction of road access to drilling sites can lead to landslides and soil erosion, especially in tropical areas with steep

hillsides and high, and occasionally intense, rainfall. The mud resulting from such erosion can choke waterways and inhibit aquatic life. Such effects can extend for large distances downstream, even as far as the coast where fishing industries may be affected.

Effects of Drilling Operations Drilling operations are generally noisy and accompanied by fumes from large diesel engines that drive the drill string and generators that provide electricity to the site. Drilling is generally a continuous (24-h, 7-day) operation, and at night powerful lights are used to illuminate the drill pad area. It may take 1–3 months to drill a deep well (1–5 km deep). The impacts of such operations on people living nearby can be severe, especially at night.

Disposal of Waste Drilling Fluid Drilling involves using a thixotropic fluid (“mud”) to provide hydrostatic pressure to prevent formation fluids from entering into the well bore, cooling the drill bit, bringing up drill cuttings, and suspending the drill cuttings, while drilling is paused and the drill string is brought in and out of the hole [5]. The mud is generally a mixture of water and clays such as bentonite, together with additives such as barium sulfate (barite), natural and synthetic polymer, asphalt and gilsonite. Other chemicals (e.g., Potassium formate) are often added to achieve various effects such as controlling viscosity, shale stability, enhancing the drilling rate, and cooling and lubricating the equipment. Deflocculants, such as anionic polyelectrolytes (e.g., acrylates, polyphosphates, lignosulfonates, or tannic acid derivatives), are frequently used to reduce viscosity of the drilling fluid. If drilling encounters a highly permeable formation and fluid is lost then bridging agents such as calcium carbonate or ground cellulose are added. If some of these chemical agents are released into natural waterways they can kill aquatic life and result in groundwater contamination.

Mass Withdrawal

Large-scale exploitation of liquid-dominated, high-temperature geothermal systems involves the withdrawal of large volumes (and hence mass) of geothermal fluid from the ground which, if not corrected for, may cause:

Pressure Declines in the Reservoir Withdrawal of large volumes of geothermal fluid are associated mainly with electric power generation. After passing through the power plant the fluid withdrawn (condensed steam and separated water) is usually injected back into the ground, however, the injection wells are generally located away from the production wells to reduce the chances of the cooler-injected water returning to the production wells and reducing the temperature of production fluids. Even if all the waste liquid is injected, there may be mass loss (up to 30% of that withdrawn) associated with evaporation of condensate from ponds where the water is cooled before injection and from the discharge of non-condensable gases into the atmosphere from the power station. A major consequence of the mass loss from parts of the field is often the formation of a two-phase (steam + water) zone in the upper part of the reservoir in the vicinity of the production wells, and as production continues this zone increases in size and the pressures decrease (both in and below this zone). At Wairakei (New Zealand), the deep (liquid phase) pressures declined by about 0.3 MPa (three bar) during discharges associated with exploratory drilling, and a further 2 MPa (20 bar) during the first 10 years of production, although subsequent pressure declines were less than 0.5 MPa (five bar) and pressures have risen since 1997 [6]. Pressure declines in the reservoir, as a result of mass withdrawal and net mass loss, are an important cause of environmental changes at or near the surface.

Degradation of Thermal Features In their natural, unexploited state many high-temperature geothermal systems are manifested at the surface by thermal features such as geysers, fumaroles, hot springs, hot pools, mud pools, sinter terraces, and thermal ground with special plant species (Figs. 1–4). Often these features are of great cultural significance [3], as well as being important tourist attractions. The thermal features result from the (upward) leakage of boiling geothermal fluid from the upper part of the reservoir, through overlying cold groundwater, to the surface.

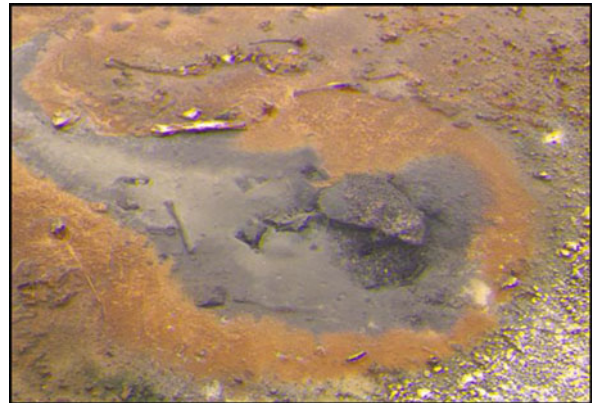
Historical evidence shows that natural thermal features have been affected, often severely, during the development and initial production stages of most high-temperature geothermal systems. At Wairakei (New Zealand), nearly all the thermal features in the Waioara and Geyser Valleys (including more than 20 geysers)



Geothermal Resources, Environmental Aspects of.

Figure 1

Beehive Geyser, Yellowstone National Park, USA



Geothermal Resources, Environmental Aspects of.

Figure 2

Hot spring, Norris Geyser Basin, Yellowstone National Park, USA

have died or been significantly reduced (Fig. 5, [7]). At Ohaaki (New Zealand), the level and temperature of water in the Ohaaki Pool declined soon after exploration drilling and reservoir testing began, but has been restored by discharging some of the warm waste water from the power plant into the pool [8]. At Tongonan field (Philippines) flow from hot springs decreased after production began [9]. Such effects are not confined to liquid-dominated systems. At Larderello (Italy), where the original natural activity consisted of numerous steam and gas jets, activity has now largely ceased. At The Geysers (USA), there has been a decrease in the flow from hot springs since exploitation began.



Geothermal Resources, Environmental Aspects of.

Figure 3

Morning Glory hot pool, Yellowstone National Park, USA



Geothermal Resources, Environmental Aspects of.

Figure 4

Minerva Terraces, Yellowstone National Park, USA

The decline in thermal features appears to be associated with a decline in reservoir pressure. As the pressure declines, so also does the amount of geothermal fluid reaching the surface and hence the thermal features decline in size and vigor. If pressures fall further then the features may die and the flow may reverse with cold groundwater flowing down into the reservoir; once this situation has occurred it may take a long time to resurrect the features. To reduce the possibility of this occurring, injection is undertaken to keep reservoir pressures as high as possible.

Depletion of Groundwater Most high-temperature geothermal systems are overlain by a cold groundwater

zone. If exploitation of the system results in a large pressure drop in the reservoir, this groundwater may be drawn down into the upper part of the reservoir in places where there are suitable high-permeability paths (such as faults). If the lateral permeability of the rocks in the groundwater zone is low then a downflow may result in a drop in the groundwater level. For example, at Wairakei, a localized drop of more than 30 m in groundwater level has occurred associated with an area of cold downflow (Fig. 6, [10]).

Downflows, and resultant groundwater level changes, may also occur as a result of breaks in the casing of disused wells [11]. Such downflows may have flow rates of up to 100–150 t/h.

Ground Deformation Withdrawal of fluid from an underground reservoir can result in a reduction of formation pore pressure which may lead to compaction in rock formations having high compressibility and result in subsidence and horizontal movements at the surface. Such ground movements have also been observed in groundwater reservoirs [12, 13] and petroleum reservoirs [14, 15], and can have serious consequences for the stability of pipelines, drains, and well casings. If the geothermal field is close to a populated area, then subsidence could lead to instability in dwellings and other buildings.

The largest recorded subsidence in a geothermal field (15 m) is in part of the Wairakei-Tauhara field (New Zealand) [16]. Here the subsidence has caused: compressional and tensional strain on pipelines and lined canals, deformation and breaking of drill casing, tilting of buildings and the equipment inside, breaking of sealed road surfaces and alteration of the gradient of streams (Fig. 7, [17]). However, the greatest subsidence rates are confined to small areas (called “bowls”) and have decreased since reservoir pressures have stabilized (Fig. 8). In the central part of the bowl there is compressional deformation that may manifest itself as pressure ridges, and on the edges there is tensional deformation that may manifest itself as cracks in the ground (Fig. 9, [16]).

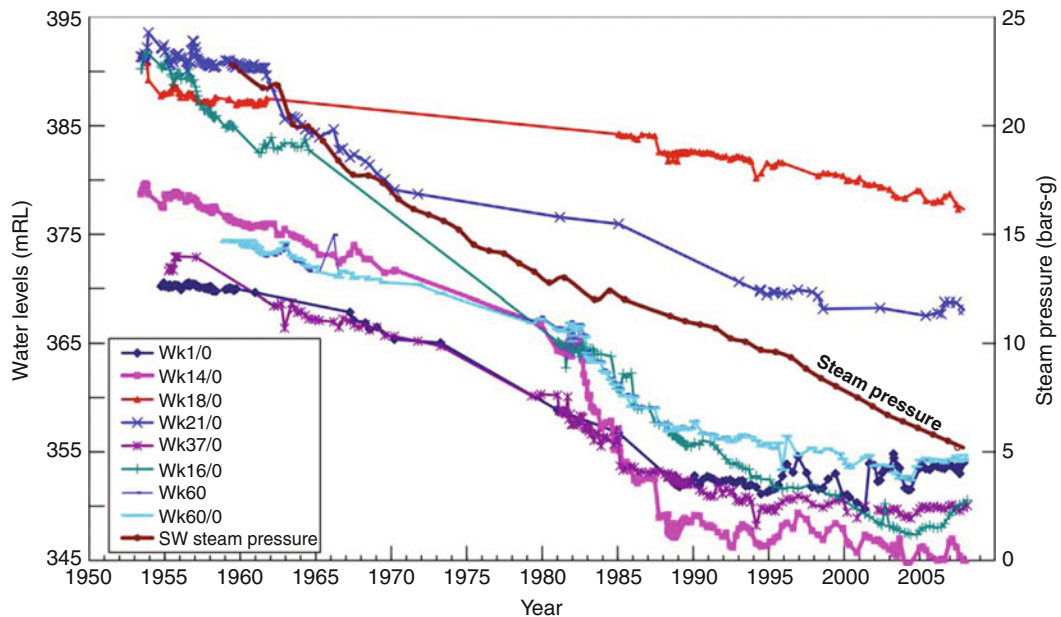
Ground movements have been recorded in other high-temperature geothermal fields in New Zealand (Table 1), at Cerro Prieto (Mexico) [18], Larderello (Italy) [19], and The Geysers (USA) [20, 21].

Ground Temperature Changes The formation and expansion of a two-phase zone, in the early stages of



Geothermal Resources, Environmental Aspects of. Figure 5

Champagne Cauldron in Geyser valley at Wairakei, New Zealand, before (*left*, painting by T. Ryan) and after 20 years of development (*right*). Arrows point to the same rock promontory



Geothermal Resources, Environmental Aspects of. Figure 6

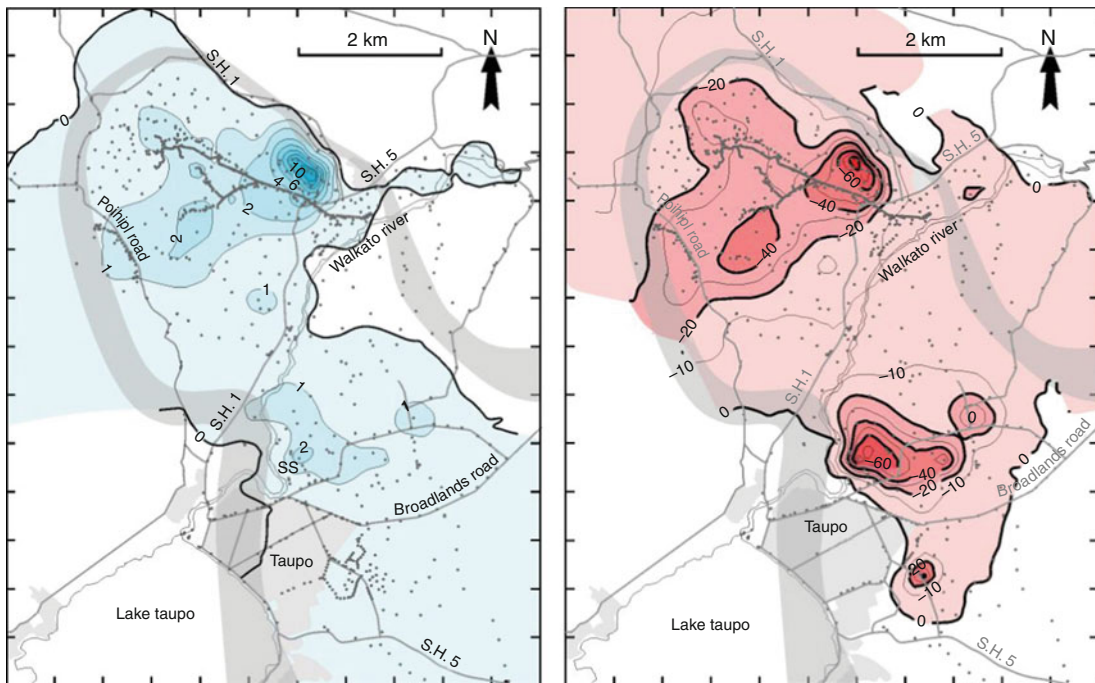
Changes in level of groundwater in shallow monitor holes at Wairakei (Taken from [10])



Geothermal Resources, Environmental Aspects of. **Figure 7**

Ponding of the Wairakei Stream near the center of the main subsidence bowl at Wairakei, New Zealand, due to change in the gradient of the stream bed

exploitation of a liquid-dominated geothermal system, can also alter the natural surface heat flow (heat loss). Steam is much more mobile than water; it can move through small fractures that are impervious to water and can move much more quickly through larger fractures. The generation and movement of steam can therefore result in increased heat flow and increased ground temperatures, so that some vegetation may become stressed or killed, generally to be replaced by other, thermally tolerant, species. At Wairakei, heat loss from natural thermal features was about 400 MW_{thermal} prior to the start of exploitation in 1958, increased to a peak of nearly 800 MW_{thermal} by the mid-1960s, and has since declined to about 600 MW_{thermal} (Fig. 10, [27]). Most of this increase was associated with increased thermal activity in the Karapiti Thermal Area, which is situated 3 km south-west of the



Geothermal Resources, Environmental Aspects of. **Figure 8**

Maps of ground subsidence and subsidence rate at Wairakei-Tauhara geothermal field, New Zealand. Left hand map shows total subsidence (m) for the period 1953–2005. Right hand map shows subsidence rate (mm/year) for the period 2001–2005. Gray zone indicates electrical resistivity boundary of the field and dots indicate measurement points (Taken from [16]). Note that the greatest subsidence rates are confined to several small parts of the field



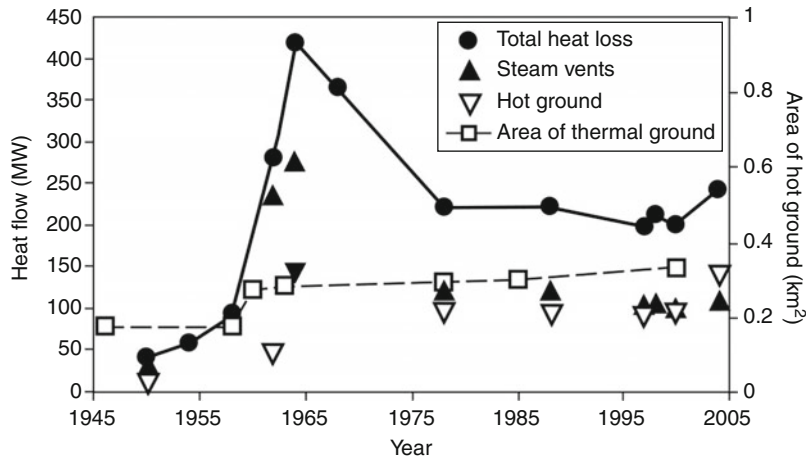
Geothermal Resources, Environmental Aspects of. Figure 9

Compressional (*left*) and tensional (*right*) deformation features at Ohaaki geothermal field, New Zealand

Geothermal Resources, Environmental Aspects of. Table 1 Ground subsidence in some producing geothermal fields. Survey dates are the period for which subsidence has been reported; it is not necessarily the whole production period. Note also that the subsidence rate changes with time and location

Field	Country	Survey period	Max. subsidence (m)	Max. rate (mm/year)	Mass change during survey period		Reference
					Withdrawal (Mt/year)	Reinjection (Mt/year)	
Wairakei	NZ	1955–1995	13.5	470	10–74	0	[16]
Ohaaki	NZ	1988–1998	1.3	400	14–17	10–13	[17]
Kawerau	NZ	1970–1996	0.48 ^a	30	10	1.7–2.6	[22]
The Geysers	USA	1977–1996	0.90	47	70	21	[20]
Bulalo	Philippines	1980–1999	0.57	32			[23]
Travale	Italy	1973–1991	0.4	25	2.3–3.9		[24]
Takigami	Japan	1992–1998	<0.02	-	11	8.8	[25]
Hatchobaru	Japan	1990–1996	0.015	-	17.7	20	
Cerro Prieto	Mexico	1994–1997	0.5	120	100	1	[18]
Svartsengi	Iceland	1976–99	0.24	14	0–9		[26]

^aCorrected for vertical displacement caused by the March 1987 Edgumbe earthquake



Geothermal Resources, Environmental Aspects of. Figure 10

Changes in heat flow with time at Karapiti thermal area, Wairakei, New Zealand. Note the transient increase in heat flow from steam vents in the early 1960s soon after production began, followed by a decrease and stabilization as steam pressures declined. Figure taken from [27]

main production borefield. The increase has been attributed to steam, associated with lateral expansion of the steam zone resulting from pressure decreases, rising to the surface up narrow fissures that were previously impervious to water. However, the increase in heat flow has resulted in an increased number of fumaroles, which has enhanced the visitor experience of tourists to the area.

Waste Liquid Disposal

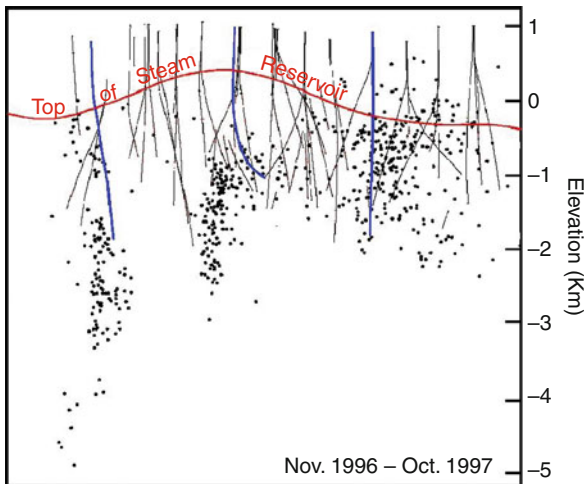
Most geothermal energy developments bring fluids containing dissolved minerals to the surface in order to extract some of the heat contained within them. In high-temperature liquid-dominated geothermal fields, the volumes of resultant liquid waste involved may be large: at Wairakei-Tauhara, a medium-sized geothermal power station (156 MW), it is currently about 5,800 m³/h. For vapor-dominated systems it is less, and for low-temperature systems it is usually much less (e.g., at Chevilly-Larue (France) it is only about 3 m³/h). After extracting some of the heat, the waste water is generally disposed of by reinjecting it deep into the ground. Surface disposal, such as putting it into waterways or evaporation ponds, may cause more environmental problems than injection because of the dissolved minerals, particularly arsenic, mercury, and boron compounds. The best method of disposal

depends on the chemistry of the geothermal fluid. For some high-temperature power plants such as at Nesjavellir (Iceland) the waste water is piped to the city of Reykjavik for district heating and then disposed of into shallow groundwater aquifers, and tests have shown no apparent effects on water chemistry. At Svartsengi power plant the waste water is discharged into the Blue Lagoon, which is a famous tourist attraction in Iceland (Fig. 11). In many cases, the waste water can be used for greenhouse heating, warm water aquaculture, space heating, irrigation swimming pools, and spas before being disposed of at the surface.

Induced Seismicity Most high-temperature geothermal systems lie in tectonically active regions where there are high levels of stress in the upper parts of the crust; this stress is manifested by active faulting and numerous earthquakes resulting from sudden relief of this stress. Studies in many high-temperature geothermal fields have shown that exploitation can result in an increase (above the normal background) in the number of small magnitude earthquakes (micro-earthquakes) within the field [28–30] (Figs. 12 and 13). Induced seismicity occurs in high-temperature fields (both liquid- and vapor-dominated), but has not been observed in low-temperature fields tapping shallow aquifers. Induced seismicity also occurs in Enhanced



Geothermal Resources, Environmental Aspects of. Figure 11
Svartsengi power plant (*left*) and the Blue Lagoon (*right*)



Geothermal Resources, Environmental Aspects of. Figure 12

Cross section through The Geysers field (California, USA) showing locations of earthquakes (*black dots*) during a 12-month period. Injection wells are shown in blue. Earthquake hypocenters and wells within 2000 ft (610 m) of the section line have been projected onto the cross section. Note that the earthquakes may extend to depths greatly below the bottom of some injection wells (Taken from [29])

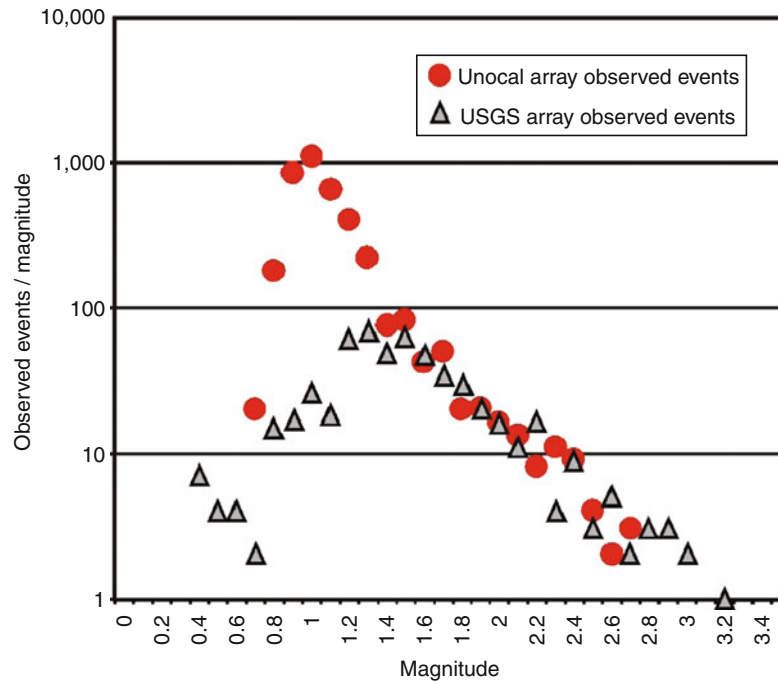
Geothermal Systems (Fig. 14) where high wellhead injection pressures are used to stimulate production by creating new fractures and extending existing fractures [31]. To date no serious damage has been caused by such earthquakes, but they may result in

a temporary shutdown of the power plant. However, the earthquakes can frighten people (especially those in non-tectonic areas who are unaccustomed to earthquakes); this has happened at Soultz in France, Basel in Switzerland, and Landau in Germany.

It is believed the increase is caused mainly by injection because when injection is stopped the number of small earthquakes decreases, and when it is restarted the number increases [28, 31]. High wellhead injection pressures increase the pore pressure at depth, particularly in existing fractures, which allows movement to suddenly release the stress and generate an earthquake. However, there are other, secondary mechanisms for induced earthquakes. Cool injected water can cause contraction of fracture surfaces leading to slight opening of the fractures, reducing static friction and triggering slip on a fracture already near failure. Also, volume changes associated with production and injection may cause perturbation in local stress conditions leading to seismic slip in a similar manner to “rockbursts” in mines as the surrounding rock adjusts to newly created void spaces [31].

Induced earthquakes may number several thousand per year in a geothermal field but few are felt, although a small number of earthquakes may reach (Richter) Magnitude 4. Detailed studies show that the induced micro-earthquakes cluster (in space) around and below the bottom of injection wells, and so the effects at the surface are generally confined to the field [32, 33].

During injection to improve an EGS reservoir at a depth of about 5 km beneath the Swiss city of Basel,



Geothermal Resources, Environmental Aspects of. Figure 13

Plot of frequency against magnitude for microearthquakes in The Geysers field, California, USA for a two-year period (1996–1998). Figure taken from [29]. Note the logarithmic increase in numbers of events as magnitude decreases. The apparent decrease in number of events below about $M = 1.3$ for the USGS array is because detection of all very small magnitude events ($M < 1$) was not possible with the equipment used, i.e., $M = 1.3$ is the effective recording threshold. Similarly for the Unocal array the effective recording threshold is about $M = 1.0$.

an earthquake of magnitude 3.4 was triggered on 8 December 2006. This event caused damage to property, and the developer's insurance paid out damages of about 7 million CHF. Recent probabilistic modeling of the seismic risk [34] (IGA, 2010) has indicated the likelihood of 40 million CHF of property damage and a 15% probability that damages could exceed 600 million CHF if development continued. While the risk of the geothermal project to cause bodily harm is low, that for property damage is considered unacceptable according to Swiss risk criteria and the development has stopped. However, other locations in Switzerland offer a significantly lower seismic risk. The incident at Basel has highlighted the need for thorough evaluation of site-specific seismic risk for future geothermal developments, especially EGS projects in densely inhabited areas.

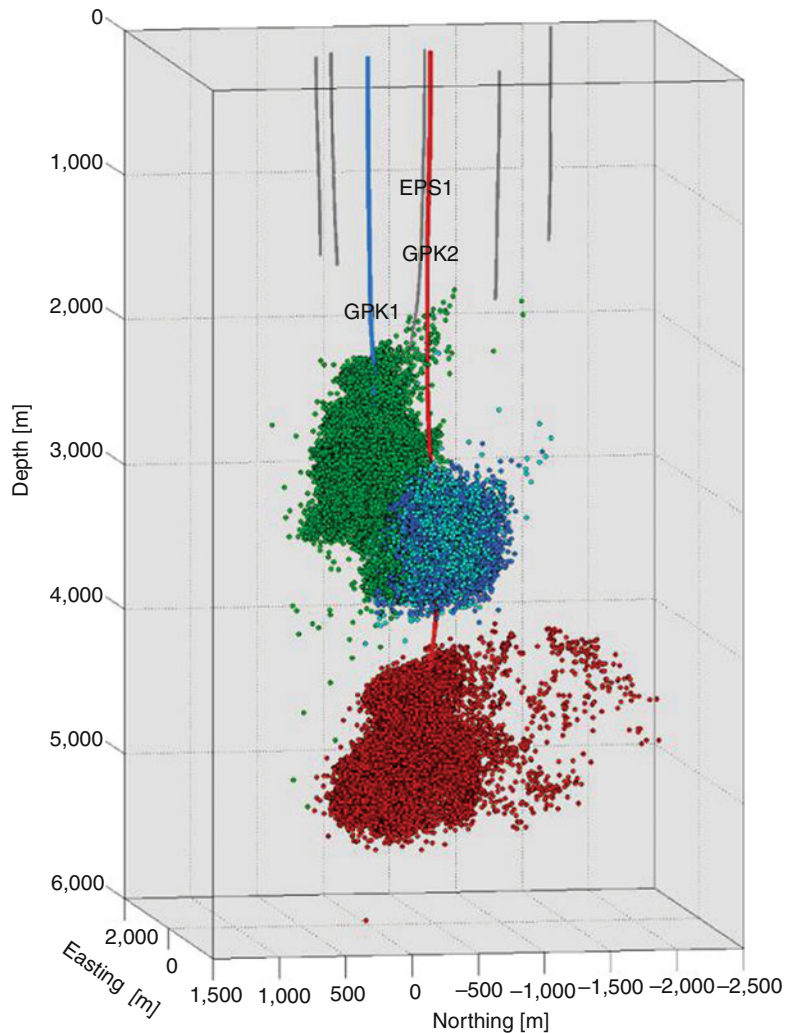
Effects on Living Organisms If hot geothermal waste water from a standard steam-cycle power station

is released directly into an existing natural waterway the localized increase in water temperature may kill fish and aquatic plants near the outlet; therefore, the water is generally passed through cooling ponds and mixed with cold water before being released. Release of untreated mineralized water into a waterway can result in chemical poisoning of fish, and also birds and animals that reside near the water because some of the toxic substances move up the "food chain."

Contamination of Groundwater Release of waste water into cooling ponds or waterways may result in shallow groundwater supplies becoming contaminated and unfit for human use, so care must be taken that the sides and bottoms of such ponds are sealed.

Waste Gas Disposal

Gas discharges from low-temperature systems are minor and do not usually cause significant environmental



Geothermal Resources, Environmental Aspects of. Figure 14

Clusters of earthquakes resulting from different injection tests at the Soultz EGS, France. Note how the earthquakes cluster around the bottom of the injection wells

impacts. In high-temperature geothermal fields, power generation using a standard steam-cycle or hybrid plant may result in the release of non-condensable gases (NCG) and fine solid particles (particulates) into the atmosphere [35]. In vapor-dominated fields where waste fluids are injected, non-condensable gases in steam will be the most important discharges from an environmental perspective. The NCG emissions are mainly from the gas exhausters of the power station, often discharged through a cooling tower. Gas and particulate discharges during well drilling, bleeding, cleanouts and testing, and from line valves and waste bore water degassing

are usually insignificant. The concentration of NCG varies not only between fields but can also vary from well to well within a field, thus changes to the proportion of steam from different wells may cause changes in the amounts of NCG discharged.

Gas concentrations and compositions cover a wide range, but the predominant gases are carbon dioxide (CO_2) and hydrogen sulfide (H_2S), together with small amounts of ammonia, mercury and boron vapor, and hydrocarbons such as methane. In some fields, up to 30% (by weight) of the geothermal fluid is NCG. However, although these amounts appear large they are

relatively small compared with fossil fuel power stations: the amount of H_2S emitted from a geothermal power station (average 0.03 g/kWh) is less than 2% of that from equivalent size coal- and oil-fired power stations (9.23 and 4.95 g/kWh, respectively). In high-temperature geothermal fields, measured direct CO_2 emission from the operation of conventional power or heating plants is widely variable, ranging from 0 to 740 g/kWh, but averages about 120 g/kWh (weighted average of 85% of the world geothermal power plant capacity) [36]. This is much less than that of 915 g/kWh from a coal-fired plant (35% efficiency), 760 g/kWh from an oil-fired plant (35% efficiency), or 315 g/kWh from a combined cycle gas plant (60% efficiency). In Enhanced Geothermal Systems, power plants are likely to be designed as closed-loop circulation systems, with no direct emissions; only if boiling occurs within the loop may some NCG emission occur. Geothermal power plants are also environmentally friendly with regard to the minor NCG: a coal-fired power plant produces the following kilograms of emissions per MWh as compared to a geothermal power plant: 4.71 versus up to 0.16 for sulfur dioxide, 1.95 versus 0 for nitrogen oxides, and 1.01 versus 0 for particulate matter [37]. Hydrogen sulfide is routinely treated at geothermal power plants, and converted to elemental sulfur.

In low-temperature fields ($<100^\circ\text{C}$), direct CO_2 emission from geothermal fluid is about 0–1 g/kWh depending on the carbonate content of the water. If the extracted geothermal fluid is passed through a heat exchanger and then completely re-injected (such as in a closed-loop pumped system), then CO_2 emission is nil to negligible.

Carbon dioxide occurs in all geothermal fluids but is most prevalent in fields in which the reservoir contains sedimentary rocks, and particularly those with limestones. Carbon dioxide is generally the most abundant NCG ($>90\%$). It is colorless and odorless, and is heavier than air so it can accumulate in topographic depressions where there is still air. It is not highly toxic (c.f. hydrogen sulfide) but at high concentrations can be fatal due to alteration of pH in the blood. There is some evidence that in high-temperature fields, the amount of CO_2 discharged (per unit mass withdrawn) decreases with time as a result of de-gassing of the deep reservoir fluid. When examining the CO_2 emissions

from geothermal power plants it is necessary to consider what would be emanating from the ground naturally (e.g., from fumaroles) in the vicinity of the plant. A strong case can be made for subtracting the natural background emission rate pre-development from the rate being released by the operation of the geothermal development. This is particularly relevant to the Larderello field in Italy where there has been a noticeable and measurable decrease in the natural release of CO_2 from the ground as a result of the geothermal power development on the field [36].

The main effects of release of non-condensable gases, together with water vapor from the cooling circuits, are local microclimatic effects such as fog. However, CO_2 need not be released directly into the atmosphere. It may be captured, purified of other gases (especially H_2S), and used to enhance plant growth in greenhouses growing vegetables. Studies have shown that as CO_2 concentration is increased from a normal level of 300 ppm to levels of approximately 1,000 ppm, crop yields may increase by up to 15% [38]. Another use of geothermal CO_2 is in carbonated drinks – at Kizildere power plant, Turkey, the NCG is scrubbed of H_2S , and the CO_2 recovered provides around 80% of the CO_2 used by the country's soft drinks industry.

Catastrophic Events

Hydrothermal Eruptions Hydrothermal eruptions have occurred at several high-temperature geothermal fields. These are small, shallow-sourced, steam and soil eruptions that generally result in craters 10–50 m in diameter and 5–20 m deep (Fig. 15). Material ejected from the craters may be deposited up to several hundred meters away. At present they cannot be reliably predicted, however, several causes have been identified that increase the likelihood of an eruption [39]. One mechanism assumes an expanding two-phase zone in the reservoir (due to production) that increases steam flow to the surface. Near the surface, an aquiclude may restrict the flow of steam resulting in an increase in the underlying pressures. Also, it has been noted that at Karapiti Thermal Area (New Zealand) the hydrothermal eruptions sometimes follow a long period of low rainfall. During long dry periods, the amount of water in the near-surface aquifer is reduced and further



Geothermal Resources, Environmental Aspects of.
Figure 15

Aerial view of hydrothermal eruption craters at Karapiti thermal area, Wairakei, New Zealand. The white lines are wooden boardwalks that allow tourists to walk over the area safely

increased heating and steam flow occurs. If a period of heavy rainfall then occurs, the permeability of the ground near the surface is quickly reduced by the rain, so that the steam cannot escape and pressures can increase to the point where the overlying rocks cannot contain the pressure and an eruption occurs. Another mechanism involves hydraulic fracturing, allowing a release of non-condensable gases and rapid decrease of the boiling point of hot water close to the surface. A third mechanism is a reduction in the lithostatic pressure by removal of the overburden, either naturally by landslides or by man-made excavations. There are no countermeasures available except to maintain reservoir pressures thus minimizing steam formation and concomitant increase in heat flow, and to refrain from building on or excavating in active thermal ground.

Blowouts During Drilling Drilling a deep well in a high-temperature geothermal field carries the risk of a blowout – an uncontrolled flow of underground fluid to the surface outside of the well. Although now rare, several blowouts have occurred in the past [40–42]. Common causes of blowouts are failure to adequately cement the well casing to the surrounding rock, or damage to the well casing by earth movement. Blowouts are generally brought under control by

directionally drilling a relief well from nearby to intersect the original well. This allows cement to be pumped into and around the original well to seal it.

Probably the most spectacular geothermal blowout was during drilling of well WK204 at Wairakei in 1960 [42]. A large and expanding crater, venting steam and rocks, quickly formed near the well, and it was fortunate that the drill rig and associated equipment was able to be removed before they were engulfed. All attempts to control the blowout failed. The crater grew to about 70 m in diameter and 20 m deep, then filled with boiling water that periodically geysered. This activity continued for several years and became a tourist attraction known as “the rogue bore.” However during 1973, the temperature and level of water in the crater declined and the feature dried up.

Landslides Many high-temperature geothermal fields are in mountainous regions, and geothermal wells are drilled from well pads carved out from steep slopes. There have been a few instances where landslides, triggered by heavy rainfall or inappropriate engineering works, have broken production wells leading to blowouts [42, 43] or have damaged steam pipelines. The most disastrous has been the landslide of 5 January 1991 in Zunil field, Guatemala, when 23 people were killed [43].

Land Use

Visual Impacts A geothermal plant must be located close to the resource, so there is often little flexibility in siting the plant. Geothermal plants generally have a low profile, and need not have a tall chimney such as coal- and oil-fired power plants. However, their visual impact may still be significant because high-temperature geothermal fields are often situated in areas of outstanding natural beauty and in National Parks (e.g., Japan, USA, and New Zealand). Any associated natural thermal features (e.g., geysers and hot pools) may be a tourist attraction or of historical and cultural significance. Undertaking developments in such areas may cause conflict during the processes for obtaining permits for access and to undertake drilling, and even for access to subsurface resources by directional drilling from outside such parks. Despite good examples of unobtrusive, scenically landscaped developments (e.g., Matsukawa, Japan), and

integrated tourism/energy developments (e.g., Wairakei, New Zealand and Blue Lagoon, Iceland), land use issues still seriously constrain new development options in some countries. Visual impact may be particularly high during drilling due to the presence of tall drill rigs.

Footprint A measure of optimum land use is the “footprint” occupied by geothermal installations. Taking into account surface installations (drilling pads, roads, pipelines, fluid separators, and power stations), the typical footprint of a conventional high-temperature geothermal power scheme is about 900 m²/GWh/year (for 30 years), or 160 m²/GWh/year excluding wells. Low-temperature geothermal plants (excluding wells) would require land use of between 1,400 and 2,300 m²/MW [44] equivalent to 150–300 m²/GWh per year. Subsurface geothermal resources accessed by directional or vertical boreholes typically occupy an area equivalent to about 10 MW/km². Therefore, about 95% of the land above a typical geothermal resource is not needed for surface installations, and can be used for other purposes (e.g., farming, horticulture, and forestry at Mokai and Rotokawa fields in New Zealand, and a game reserve at Olkaria, Kenya).

Methods of Avoiding or Minimizing Impacts

Good Management Practices

Responsibility for protecting the environment rests with the developer, and specifically with the engineers and managers. Some general principles for protecting the environment are: regular monitoring of the environment; reliance on scientists and engineers to recognize problems; acting before scientific consensus is achieved; confronting uncertainty; including human motivation (short-sightedness and greed); taking a precautionary approach and being prepared for worst-case scenarios. These principles can be encouraged by regulators which provide good environmental strategies (such as avoid, remedy, or mitigate) and sound guidelines.

Good Engineering Practices

Site Investigation and Development The impacts of access roads and field development can be minimized by careful planning and construction of access roads and prompt re-planting of vegetation destroyed. Simple and

reliable criteria have been developed to assess slope instability and the potential for slope failure in the Philippines [45], and here remediation includes construction of benches to prevent landslides and rigid structural barriers up-slope of and over pipework.

Noise Reduction Noise inevitably occurs during the exploration drilling, construction, and production phases of development. Air drilling is the noisiest (120 dBA) due to the “blow pipe effect” where the gases exit, but suitable muffling can reduce this to around 85 dBA [46]. Mud drilling is quieter at around 80 dBA. Diesel engines operating compressors and electricity generators can also produce a low-frequency resonant sound that carries for long distances; this noise can be constrained, to less than 55 dBA during the day and 45 dBA at night, by suitable muffling and confining noisy operations (such as tripping or cementing) to the daytime hours. Construction of screens of sound-absorbing material, such as vegetation, is also used to reduce the impacts of drilling noise. Following drilling, a well is usually discharged to remove drilling debris. Such vertical discharges are very noisy (up to 120 dBA). After this, there is normally a period of well testing; this can be suitably muffled by the use of silencers, but even then the noise is still significant (70 – 110 dBA). The well is then put on “bleed” where the noise is around 85 dBA reduced to 65 dBA if the “bleed” is led to a rock muffler [46]. Drilling is generally a continuous 24-h/day operation and the effects of using powerful lamps to light the work site at night are reduced by temporary screens and careful placement of the lamps. Modern drilling techniques involve using minimal amounts of fluid and recycling as much as possible.

Injection Changes to thermal features are associated with declines in production reservoir pressures and it appears that the best way to prevent or minimize changes to these features is to minimize any reduction in reservoir pressures by undertaking injection. This will also minimize any changes in groundwater level and temperature, and avoid waste liquid contaminating groundwater. Deep injection also reduces the effects of waste liquid disposal on living organisms. However, it may be necessary to keep injection pressures to a minimum to minimize induced seismicity.

Induced seismicity may be reduced if injection is reduced or halted when the seismicity reaches

pre-determined levels. In 2003, a trial was carried out at Berlin Geothermal Field, located in a tectonically active area, which used pressurized injection in an attempt to improve fracture permeability. A calibrated real-time “traffic light” control system [47] was established to reduce or stop injection operations if the levels of vibration (peak ground velocity) from injection-induced seismicity exceeded acceptable levels (normal background = “green,” significant felt events = “orange,” and damaging events = “red”).

Engineering to Overcome Ground Movements Little can be done to prevent or minimize ground deformation, except to maintain reservoir pressures. Experience suggests that subsidence can be halted, but it is difficult to reverse by increasing reservoir pressure because of the great weight of rock overlying the formation that has compacted. The effects of deformation on pipelines is reduced by mounting the pipelines on rollers, but experience at Wairakei shows that even with such assistance sections of pipe need periodically to be removed or installed to maintain the pipeline network. Equipment that is sensitive to level is generally mounted on an adjustable base.

Plant Design Suitable design of power stations employing active monitoring systems will minimize the effects of non-condensable gas discharges and reduce microclimatic effects (e.g., suitable placement of cooling towers and gas discharge vents). If induced seismicity is likely then all structures in the field should be earthquake resistant.

Regulations

Most geothermal developments are controlled and monitored by independent regulatory authorities, such as central, regional, or local government, and they issue (to developers or users) permits or consents which ensure that the best environmental practices are followed [48–51]. This generally involves preparation of an Environmental Impact Report (EIR) before development begins; consideration of that report by officials, experts, and the public; granting of permits subject to restrictions; setting up of a monitoring program and measurements taken regularly; and periodic review of the monitoring data and renewal of the permits. This is

not entirely altruistic because if severe environmental damage occurs it is generally government that has to take ultimate responsibility for the problem.

Economic Measures

Royalties or User Charges One method of encouraging use of a geothermal resource in an efficient and sustainable manner is to charge users for the amount of energy taken. A good example of how implementation of user charges reduced wasteful practices was in Rotorua city, New Zealand [52]. Here, especially during summer, some individual well owners were taking geothermal fluid and not using it, but passing the hot water directly to waste at the surface instead of shutting down their well. Introduction of user charges, and closure of poorly performing wells and equipment, persuaded many of these owners to combine and operate a single well in a sustainable manner, with injection. Since the introduction of the charges the net amount of fluid withdrawn has decreased from about 30 kt/day to less than 10 kt/day, and water levels in the shallow thermal aquifers have risen.

Bonds Another economic measure that can be effectively used to protect the environment and encourage sustainable development is the requirement for a developer to deposit a large refundable bond that is forfeited if environmental damage occurs. Interest on the bond money, less an amount to cover taxes and inflation, may be returned annually to the developer. Although the damage may not be able to be rectified by money, the potential loss of a large amount of money may keep a company more focussed on the environment and the consequences of its actions. Such a system is effective when there is the suspicion that a development company will not be able to meet its obligations either through lack of expertise or financial problems. Bonds are particularly effective with public companies where the profits, share value, and bonuses of the managers may be adversely affected by loss of the bond. However, to date, this approach has not been used in the geothermal industry; it is more common in the mining industry where developers are transient.

Future Directions

The worldwide need for sustainable energy sources means that the exploitation of geothermal energy will

continue and even accelerate in the next few decades. Although a few geothermal developments have had serious environmental effects, such as at Wairakei, most of these cases were before the dynamics of geothermal systems was understood. The causes of most of the environmental effects have since been recognized and countermeasures are employed as part of permitting. International scientific efforts, under the auspices of the International Energy Agency (IEA), to devise methods of extracting geothermal energy with the minimum of environmental effects were started in the late 1990s and will continue.

Bibliography

Primary Literature

- Parker SP (1980) Encyclopedia of environmental science, 2nd edn. McGraw-Hill, New York
- Brown L (1993) New shorter oxford english dictionary. Clarendon, Oxford
- Cataldi R, Hodgson SF, Lund JW (1999) Stories from a heated Earth: our geothermal heritage. Geothermal Resources Council, International Geothermal Association, Sacramento, p 569
- Thoreaux HD (1860) Letter to Harrison Blake, Written 20 May, 1860. www.walden.org/documents/file/Library/Thoreau/writings/correspondence/Correspondence/1860. Accessed 14 April 2011
- DiPippo R (2008) Geothermal well drilling. In: DiPippo R (ed) Geothermal power plants, 2nd edn. Elsevier, Amsterdam, pp 39–48
- Bixley PF, Clotworthy AW, Mannington WI (2009) Evolution of the Wairakei geothermal reservoir during 50 years of production. *Geotherm* 38:145–154
- White PA, Hunt TM (2005) Simple modelling of the effects of exploitation on hot springs, Geyser Valley, Wairakei, New Zealand. *Geotherm* 34:184–204
- Glover RB, Hunt TM, Severne CM (2000) Impacts of development on a natural thermal feature and their mitigation – Ohaaki Pool, New Zealand. *Geotherm* 29:509–523
- Bolanos GT, Parilla EV (2000) Response of the Bao-Banati thermal area to development of the Tongonan geothermal field, Philippines. *Geotherm* 29:499–508
- Bromley CJ (2009) Groundwater changes in the Wairakei-Tauhara geothermal system. *Geotherm* 38:134–144
- Bixley PF, Hattersley SD (1983) Long term casing performance of Wairakei production wells. In: Proceedings 5th NZ geothermal workshop, Geothermal Institute, Auckland, pp 257–263
- Lofgren BE, Klausing RL (1969) Land subsidence due to ground-water withdrawal, Tulare-Wasco area, California. US geological survey professional Paper 437-B
- Holzer TL, Johnson AI (1985) Land subsidence caused by ground water withdrawal in urban areas. *GeoJ* 11:245–255
- Mayuga MN, Allen DR (1970) Subsidence in the Wilmington oil field, Long Beach, California, U.S.A. In: Tison LJ (ed) Land subsidence. International Association of Scientific Hydrology, UNESCO, pp 66–79
- Bruno MS, Bovberg CA (1992) Reservoir compaction and surface subsidence above the Lost Hills Field, California. 33rd US symposium on rock mechanics, Santa Fe, New Mexico, USA, Paper No. 92-0263
- Allis RG, Bromley CJ, Currie S (2009) Update on subsidence at the Wairakei-Tauhara geothermal system, New Zealand. *Geotherm* 38:169–180
- Allis RG, Zhan X (2000) Predicting subsidence at Wairakei and Ohaaki geothermal fields, New Zealand. *Geotherm* 29:479–497
- Glowacka E, Sarychikhina O, Nava FA (2005) Subsidence and stress change in the Cerro Prieto geothermal field, B.C., Mexico. *Pure Appl Geophys* 162:2095–2110
- Geri G, Marson I, Rossi A, Toro B (1982) Gravity and elevation changes in the Travale geothermal field (Tuscany) Italy. *Geotherm* 11:153–161
- Mossop A, Segall P (1997) Subsidence at the Geysers geothermal field, N. California from a comparison of GPS and leveling surveys. *Geophys Res Lett* 24:1839–1842
- Gettings P, Allis R, Harris RN, Chapman DS (2001) High-precision gravity and GPS monitoring of The Geysers geothermal system. American Geophysical Union, Fall Meeting, Abstract #G41B-0225
- Allis RG, Carey B, Darby D, Read SAL, Rosenberg M, Wood CP (1997) Subsidence at Ohaaki field. In: Proceedings of the 19th NZ geothermal workshop, Geothermal Institute, Auckland, pp 9–15
- Protacio JA, Golla G, Nordquist G, Acuña J, San Andres RB (2000) Gravity and elevation changes in the Bulalo geothermal field, Philippines: independent checks and constraints on numerical simulation. In: Proceedings of the 22nd NZ Geothermal Workshop, Geothermal Institute, Auckland, pp 115–119
- Di Filippo M, Dini I, Marson I, Palmieri F, Rossi A, Toro B (1995) Subsidence and gravity changes induced by exploitation in the Travale-Radicondoli geothermal field (Tuscany, Italy). In: Proceedings world geothermal congress 1995, Florence, Italy, 3, pp 1945–1949
- Fujimitsu Y, Nishijima J, Shimosako N, Ehara S, Ikeda K (2000) Reservoir monitoring by repeat gravity measurements at the Takigami geothermal field, central Kyushu, Japan. In: Proceedings world geothermal congress 2000, Japan, pp 573–577
- Eysteinnsson H (2000) Elevation and gravity changes at geothermal fields on the Reykjanes peninsula, SW Iceland. In: Proceedings world geothermal congress 2000, Kyushu – Tohoku, pp 559–564
- Hunt TM, Bromley CJ, Risk GF, Sherburn S, Soengkono S (2009) Geophysical investigations of the Wairakei field. *Geotherm* 38:85–97
- Sherburn S, Allis RG, Clotworthy A (1990) Microseismic activity at Wairakei and Ohaaki geothermal fields. In: Proceedings of the 12th NZ Geothermal Workshop, Geothermal Institute, Auckland, pp 51–55

29. Smith B, Beall J, Stark M (2000) Induced seismicity in the SE Geysers Field, California, USA. In: Proceedings world geothermal congress 2000, pp 2887–2892
30. Rivas JA, Castellón JA, Maravilla JN (2005) Seven years of reservoir seismic monitoring at Berlín geothermal field, Usulután, El Salvador. In: Proceedings world geothermal congress 2005, Antalya, Turkey, 24–29 April 2005, Paper 215 6 pp
31. Majer EL, Baria R, Stark M, Oates S, Bommer J, Smith B, Asanuma H (2007) Induced seismicity associated with enhanced geothermal systems. *Geotherm* 36:185–222
32. Stark M (1990) Imaging injected water in the Geysers reservoir using microearthquake data. *Geoth Resourc Counc Trans* 14:1697–1704
33. Baria R, Michelet S, Baumgartner J, Dyer B, Nicholls J, Hettkamp T, Teza D, Soma, N, Asanuma H, Garnish J, Megel T (2005) Creation and mapping of 5000 m deep HDR/HFR reservoir to produce electricity. In: Proceedings world geothermal congress Antalya, Turkey, paper 1627
34. IGA (2010) Seismic risk assessment in Basel. *IGA News* 79:6–9
35. Webster JG (1995) Chemical impacts of geothermal development. In: Brown KL (convenor) Course on environmental aspects of geothermal development. International Geothermal Association, Pisa, Italy, 145 pp
36. Bertani R, Thain IA (2002) Geothermal power generating plant CO₂ emission survey. *IGA News* 49:1–3. <http://www.geothermal-energy.org/files-39.html>. Accessed 13 April 2011
37. Lund JW (2007) Characteristics, development and utilization of geothermal resources. *Geo Heat Cent Q Bull* 28:1–9
38. Dunstall MG, Graeber G (1997) Geothermal carbon dioxide for use in greenhouses. *Geo Heat Cent Q Bull* 18(1):8–13
39. Bromley CJ, Mongillo MA (1994) Hydrothermal eruptions – a hazard assessment. In: Proceedings of the 16th NZ Geothermal Workshop, Geothermal Institute, Auckland, pp 45–50
40. Adams N, Thompson JD (1989) How a geothermal blowout was controlled. *World Oil* 208(6):36–42
41. Anderson I (1991) Blowout blights future of Hawaii's geothermal power. *New Sci* 1778:17
42. Bolton RS, Hunt TM, King TR, Thompson GEK (2009) Dramatic incidents during drilling at Wairakei geothermal field, New Zealand. *Geotherm* 38:40–47
43. Goff S, Goff F (1997) Environmental impacts during geothermal development: some examples from Central America. In: Proceedings NEDO international geothermal symposium, 10–14 March 1997, Sendai, Japan, pp 242–250
44. Tester JW (2006) The future of geothermal energy – impact of Enhanced Geothermal Systems (EGS) on the United States in the 21st Century. US Dept Energy, Idaho National Laboratory, Idaho Falls, USA, 372 pp. ISBN 0-615-13438-6. http://geothermal.inel.gov/publications/future_of_geothermal_energy.pdf. Accessed 13 April 2011
45. Leynes RD, Pioquinto WPC, Caranto JA (2005) Landslide hazard assessment and mitigation measures in Philippine geothermal fields. *Geotherm* 34:205–217
46. Brown KL (1995) Impacts on the physical environment. In: Environmental aspects of geothermal development. IGA pre-congress short course, world geothermal congress Pisa, Italy, May 1995, pp 39–55
47. Bommer JJ, Oates S, Cepeda JM, Lindholm C, Bird J, Torres R, Marroquín G, Rivas J (2006) Control of hazard due to seismicity induced by a hot fractured rock geothermal project. *Eng Geol* 83:287–306
48. Hietter L (1995) Introduction to geothermal development and regulatory requirements. In: Brown KL (convenor) Course notes on environmental aspects of geothermal development. International Geothermal Association, Pisa, Italy, 18–20 May 1995, 145 pp
49. Goff S (2000) The effective use of environmental impact assessments (EIAs) for geothermal development projects. In: Proceedings world geothermal congress 2000, Japan, 597–603
50. Luketina K (2000) New Zealand geothermal resource management – a regulatory perspective. In: Proceedings World Geothermal Congress 2000, Japan, pp 751–756
51. Daysh S, Chrisp M (2009) Environmental planning and consenting for Wairakei: 1953–2008. *Geotherm* 38:192–199
52. O'Shaughnessy BW (2000) Use of economic instruments in management of Rotorua geothermal field, New Zealand. *Geotherm* 29:539–556

Books and Reviews

- Dickson MH, Fanelli M (2003) Geothermal energy utilization and technology. UNESCO Publishing, Paris. ISBN 92-3-103915-6
- DiPippo R (2008) Geothermal power plants: principles, applications, case studies and environmental impact. Elsevier, Amsterdam
- Hunt TM (ed) (2000) Special issue on environmental aspects of geothermal development. *Geothermics* 29 (4/5), 175 pp
- Hunt TM (2001) Five lectures on environmental effects of geothermal utilization. Report 2000-1, United Nations University, Reykjavik, Iceland. 109 pp. ISBN-9979-68-070-9
- Hunt TM (ed) (2005) Special issue on environmental aspects of geothermal energy. *Geothermics* 34(2). ISSN 0375-6505
- Rybach L, Muffler LJP (1981) Geothermal systems: principles and case histories. Wiley, New York, 359 pp

Global Economic Impact of Transgenic/Biotech Crops (1996–2008)

GRAHAM BROOKES

Agricultural Economist, PG Economics Ltd,
Dorset, UK

Article Outline

Glossary

Definition of the Subject

Introduction

Economic Impact of Transgenic/Biotech Crops

Herbicide-Tolerant Soybeans

Herbicide-Tolerant Maize

Herbicide-Tolerant Cotton

Herbicide-Tolerant Canola

GM Herbicide-Tolerant (GM HT) Sugar Beet

GM Insect-Resistant (To Corn-Boring Pests: GM IR)
Maize

Insect-Resistant (Bt) Cotton (GM IR)

Other Biotech Crops

Indirect (Nonpecuniary) Farm-Level Economic
Impacts

Production Effects of the Technology

Bibliography

Glossary

Direct farm income benefit Improvements in income arising from changes in yield and production levels or associated with cost reductions/productivity enhancements associated with the use of transgenic crops.

Herbicide tolerance Tolerance to a herbicide (e.g., glyphosate) delivered by genetic modification techniques. This allows a crop to be sprayed with the “tolerant herbicide” without harming the crop but providing good weed control.

Insect resistance Resistance to a pest (e.g., corn-boring pests) delivered by genetic modification techniques. This allows a crop to be grown without having to use alternative methods of pest control, notably the use of insecticides.

Nonpecuniary benefit Additional farm-level benefits to direct farm income benefits that are more

intangible and difficult to measure in monetary terms (e.g., additional management flexibility).

No tillage agriculture The use of a production technique in which the soil is not tilled/plowed. It is in contrast to traditional plow-based production systems and allows farmers to save on fuel use and contributes to improved soil water retention and reduced soil erosion.

Second crop soybeans The planting of a crop of soybeans after another crop (often wheat) in the same growing season. This allows a farmer to obtain two crops from the same piece of land in one season.

Definition of the Subject

The application of biotechnology to commercial agriculture on a widespread basis has occurred since 1996. The extent of this adoption in terms of crops and (biotechnology) traits is explored and the associated economic impacts for the period 1996–2008 are assessed, to help identify some of the main reasons why farmers have adopted the technology.

Introduction

This article examines specific global socioeconomic impacts on farm income over the 13-year period 1996–2008. It also quantifies the production impact of the technology on the key crops in areas where it has been used. The analysis concentrates on farm income effects because this is a primary driver of adoption among farmers (both large commercial and small-scale subsistence). It also considers more indirect farm income or nonpecuniary benefits, and quantifies the (net) production impact of the technology. More specifically, it covers the following main issues:

- Impact on crop yields
- Effect on key costs of production, notably seed cost and crop protection expenditure
- Impact on other costs such as fuel and labor
- Effect on profitability
- Other impacts such as crop quality, scope for planting a second crop in a season and impacts that are often referred to as intangible impacts such as convenience, risk management, and husbandry flexibility
- Production effects

The contribution is based largely on extensive analysis of existing farm-level impact data for biotech crops. While primary data for impacts of commercial cultivation were not available for every crop, in every year and for each country, a substantial body of representative research and analysis is available and this has been used as the basis for the analysis presented.

As the economic performance and impact of this technology at the farm level varies widely, both between, and within regions/countries (as applies to any technology used in agriculture), the measurement of performance and impact is considered on a case-by-case basis in terms of crop and trait combinations. The analysis presented is based on the average performance and impact recorded in different crops by the studies reviewed; the average performance being the most common way in which the identified literature has reported impact. Where several pieces of relevant research (e.g., on the impact of using a GM trait on the yield of a crop in one country in a particular year) have been identified, the findings used have been largely based on the average of these findings.

This approach may both, overstate, or understate, the real impact of GM technology for some trait, crop and country combinations, especially in cases where the technology has provided yield enhancements. However, as impact data for every trait, crop, location, and year is not available, the authors have had to extrapolate available impact data from identified studies to years for which no data are available. Therefore, the authors acknowledge that this represents a weakness of the research. To reduce the possibilities of over/understating impact, the analysis:

- Directly applies impacts identified from the literature to the years that have been studied. As a result, the impacts used vary in many cases according to the findings of literature covering different years. Hence, the analysis takes into account the variation in the impact of the technology on the yield based on its effectiveness in dealing with (annual) fluctuations in pest and weed infestation levels as identified by research.
- Uses current farm-level crop prices and bases any yield impacts on (adjusted – see below) current average yields. In this way, some degree of dynamic has been introduced into the analysis that would,

otherwise, be missing if constant prices and average yields identified in year-specific studies had been used.

- Includes some changes and updates to the impact assumptions identified in the literature based on consultation with local sources (analysts, industry representatives) so as to better reflect prevailing/ changing conditions (e.g., pest and weed pressure, cost of technology).
- Adjusts downward the average base yield (in cases where GM technology has been identified as having delivered yield improvements) on which the yield enhancement has been applied. In this way, the impact on total production is not overstated.

Other aspects of the methodology used to estimate the impact on direct farm income are as follows:

- Impact is quantified at the trait and crop level, including where stacked traits are available to farmers. Where stacked traits have been used, the individual trait components were analyzed separately to ensure estimates of all traits were calculated.
- All values presented are nominal for the year shown and the base currency used is the US dollar. All financial impacts in other currencies have been converted to US dollars at prevailing annual average exchange rates for each year.
- The analysis focuses on the changes in farm income for each year, arising from the impact of GM technology on yields, key costs of production, notably seed cost and crop protection expenditure and also the impact on costs such as fuel and labor (inclusion of impact on these categories of cost are, however, more limited than the impacts on seed and crop protection costs because only a few of the papers reviewed have included consideration of such costs in their analyses). Therefore, in most cases the analysis relates to impact of crop protection and seed cost only.
- Crop quality (e.g., improvements in quality arising from less pest damage or lower levels of weed impurities that result in price premia being obtained from buyers) and the scope for facilitating the planting of a second crop in a season (e.g., second crop soybeans in Argentina following wheat that would, in the absence of the GM herbicide-tolerant

(GM HT) seed, probably not have been planted). Thus, the farm income effect measured is essentially a gross margin impact (impact on gross revenue less variable costs of production) rather than a full net cost of production assessment. Through the inclusion of yield impacts and the application of actual (average) farm prices for each year, the analysis also indirectly takes into account the possible impact of biotech crop adoption on global crop supply and world prices.

This article also examines some of the more intangible (more difficult to quantify) economic impacts of GM technology. The literature in this area is much more limited and in terms of aiming to quantify these impacts, largely restricted to the US-specific studies. The findings of this research (notably relating to the USA, and drawing on Marra and Piggot [1, 2] are summarized and extrapolated to the cumulative biotech crop planted areas in the USA over the period 1996–2008.

Lastly, this article includes estimates of the production impacts of GM technology at the crop level. These have been aggregated to provide the reader with a global perspective of the broader production impact of the technology. These impacts derive from the yield impacts (where identified), but also from the facilitation of additional cropping within a season (notably in relation to soybeans in South America).

Economic Impact of Transgenic/Biotech Crops

The section below is structured on a trait and country basis highlighting the key farm-level impacts.

Herbicide-Tolerant Soybeans

The USA

In 2008, 92% of the total US soybean crop was planted to genetically modified herbicide-tolerant cultivars (GM HT). The farm-level impact of using this technology since 1996 is summarized in [Table 1](#).

The key features are as follows:

- The primary impact has been to reduce the soybean cost of production. In the early years of adoption, these savings were between \$25/ha and \$34/ha. In recent years, estimates of the cost savings have been in the range of \$30–\$85/ha (based on a comparison

of conventional herbicide regimes in the early 2000s that would be required to deliver a comparable level of weed control to the GM HT soybean system). In 2008, the cost savings declined relative to earlier years because of the significant increase in the global price of glyphosate relative to increases in the price of other herbicides (commonly used on conventional soybeans). The main savings have come from lower herbicide costs (while there were initial cost savings in herbicide expenditure, these increased when glyphosate came off-patent in 2000. Growers of GM HT soybeans initially applied Monsanto's Roundup herbicide but over time, and with the availability of low-cost generic glyphosate alternatives, many growers switched to using these generic alternatives (the price of Roundup also fell significantly post 2000) plus a \$6–\$10/ha savings in labor and machinery costs.

- Against the background of underlying improvements in average yield levels over the 1996–2008 period (via improvements in plant breeding), the specific yield impact of the GM HT technology used up to 2008 has been neutral (some early studies of the impact of GM HT soybeans in the USA, suggested that GM HT soybeans produced lower yields than conventional soybean varieties. Where this may have occurred, it applied only in early years of adoption when the technology was not present in all leading varieties suitable for all of the main growing regions of the USA. By 1998/1999, the technology was available in leading varieties and no statistically significant average yield differences have been found between GM and conventional soybean varieties.
- The annual total national farm income benefit from using the technology rose from \$5 million in 1996 to \$1.42 billion in 2007. In 2008, the farm income was about \$1.2 billion. The cumulative farm income benefit over the 1996–2008 period (in nominal terms) was \$11 billion.
- In added value terms, the increase in farm income in recent years has been equivalent to an annual increase in production of between +5% and +10%.

Argentina

As in the USA, GM HT soybeans were first planted commercially in 1996. Since then, use of the technology

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 1 Farm-level income impact of using GM herbicide-tolerant (GM HT) soybeans in the USA 1996–2008

Year	Cost savings (\$/ha)	Net cost saving/increase in gross margins, inclusive of cost of technology (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1996	25.2	10.39	5.0	0.03
1997	25.2	10.39	33.2	0.19
1998	33.9	19.03	224.1	1.62
1999	33.9	19.03	311.9	2.5
2000	33.9	19.03	346.6	2.69
2001	73.4	58.56	1,298.5	10.11
2002	73.4	58.56	1,421.7	9.53
2003	78.5	61.19	1,574.9	9.57
2004	60.1	40.33	1,096.8	4.57
2005	69.4	44.71	1,201.4	6.87
2006	57.0	32.25	877.1	4.25
2007	85.2	60.48	1,417.2	6.01
2008	68.6	43.88	1,219.5	4.25

Sources and notes:

1. Impact data 1996–1997 based on Marra et al [3], 1998–2000 based on Carpenter and Gianessi [4] and 2001 [5] onward based on Sankala and Blumenthal [6, 7] and Johnson and Strom [8] plus updated 2008 to reflect recent changes in herbicide prices
2. Cost of technology: \$14.82/ha 1996–2002, \$17.3/ha 2003, \$19.77/ha 2004, \$24.71/ha 2005 onward
3. The higher values for the cost savings in 2001 onward reflect the methodology used by Sankala and Blumenthal, which was to examine the conventional herbicide regime that would be required to deliver the same level of weed control in a low/reduced till system to that delivered from the GM HT no/reduced till soybean system. This is a more robust methodology than some of the more simplistic alternatives used elsewhere. In earlier years, the cost savings were based on comparisons between GM HT soy growers and/or conventional herbicide regimes that were commonplace prior to commercialization in the mid-1990s when conventional tillage systems were more important

has increased rapidly and almost all soybeans grown in Argentina are GM HT (99%). Not surprisingly, the impact on farm income has been substantial, with farmers deriving important cost saving and farm income benefits both similar and additional to those obtained in the USA (Table 2). More specifically, it covers the following main issues:

- The impact on yield has been neutral (i.e., no positive or negative yield impact).
- The cost of the technology to Argentine farmers has been substantially lower than in the USA (about \$1–\$4/ha compared to \$15–\$25/ha in the USA: see Table 1) mainly because the main technology provider (Monsanto) was not able to obtain patent protection for the technology in Argentina.
- As such, Argentine farmers have been free to save and use biotech seed without paying any technology fees or royalties (on farm-saved seed) for many years and estimates of the proportion of total soybean seed used that derives from a combination of declared saved seed and uncertified seed in 2008 were about 75% (i.e., 25% of the crop was planted to certified seed).
- The savings from reduced expenditure on herbicides, fewer spray runs, and machinery use have been in the range of \$24–\$30/ha, although in 2008, savings fell back to about \$16/ha because of the significant increase in the price of glyphosate relative to other herbicides. Net income gains have been in the range of \$21–\$29/ha, although in 2008 a lower average level of about \$14/ha has occurred.

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 2 Farm-level income impact of using GM HT soybeans in Argentina 1996–2008

Year	Cost savings (\$/ha)	Net saving on costs (inclusive of cost of technology (\$/ha))	Increase in farm income at a national level (\$ millions)	Increase in farm income from facilitating additional second cropping (\$ millions)
1996	26.10	22.49	0.9	0
1997	25.32	21.71	42	25
1998	24.71	21.10	115	43
1999	24.41	20.80	152	118
2000	24.31	20.70	205	143
2001	24.31	20.70	250	273
2002	29.00	27.82	372	373
2003	29.00	27.75	400	416
2004	30.00	28.77	436	678
2005	30.20	28.96	471	527
2006	28.72	26.22	465	699
2007	28.61	26.11	429	1,134
2008	16.37	13.87	233	765

Sources and notes:

1. The primary source of information for impact on the costs of production is Qaim and Traxler [9, 10]. This has been updated in recent years to reflect changes in herbicide prices
2. All values for prices and costs denominated in Argentine pesos have been converted to US dollars at the annual average exchange rate in each year
3. The second cropping benefits are based on the gross margin derived from second crop soybeans multiplied by the total area of second crop soybeans (less an assumed area of second crop soybeans that equals the second crop area in 1996 – this was discontinued from 2004 because of the importance farmers attach to the GM HT system in facilitating them remaining in no tillage production systems). The source of gross margin data comes from Grupo CEO
4. Additional information is available in [Appendix 1](#)
5. The net savings to costs understate the total gains in recent years because two thirds to 80% of GM HT plantings have been to farm-saved seed on which no seed premium was payable (relative to the \$3-\$4/ha premium charged for new seed)

- The price received by farmers for GM HT soybeans in the early years of adoption was, on average, marginally higher than for conventionally produced soybeans because of lower levels of weed material and impurities in the crop. This quality premia was equivalent to about 0.5% of the baseline price for soybeans.
- The net income gain from the use of the GM HT technology at a national level was \$233 million in 2008. Since 1996, the cumulative benefit (in nominal terms) has been \$3.57 billion.
- An additional farm income benefit that many Argentine soybean growers have derived comes from the additional scope for second cropping of soybeans.

This has arisen because of the simplicity, ease, and weed management flexibility provided by the (GM) technology, which has been an important factor facilitating the use of no and reduced tillage production systems. In turn, the adoption of low/no tillage production systems has reduced the time required for harvesting and drilling subsequent crops and hence has enabled many Argentine farmers to cultivate two crops (wheat followed by soybeans) in one season. As such, 20% of the total Argentine soybean crop was second crop in 2008 (3.4 million hectares), compared to 8% in 1996. Based on the additional gross margin income derived from second crop soybeans (see [Appendix 1](#)), this has contributed a further boost

to national soybean farm income of \$765 billion in 2008 and \$5.19 billion cumulatively since 1996.

- The total farm income benefit inclusive of the second cropping was \$998 million in 2008 and \$8.76 billion cumulatively between 1996 and 2008.
- In added value terms, the increase in farm income from the direct use of the GM HT technology (i.e., excluding the second crop benefits) in the last 3 years has been equivalent to an annual increase in production of between +2% and +7%. The additional production from second soybean cropping facilitated by the technology in 2008 was equal to 20% of total output.

Brazil

GM HT soybeans were probably first planted in Brazil in 1997. Since then, the area planted has increased to

62% of the total crop in 2008 (until 2003 all plantings were technically illegal).

The impact of using GM HT soybeans has been similar to that identified in the USA and Argentina. The net savings on herbicide costs have been larger in Brazil due to higher average costs of weed control. Hence, the average cost saving arising from a combination of reduced herbicide use, fewer spray runs, labor and machinery savings were between \$30/ha and \$81/ha in the period 2003–2008 (Table 3). The net cost saving after deduction of the technology fee (assumed to be about \$20/ha in 2008) has been between \$9/ha and \$61/ha in recent years. At a national level, the adoption of GM HT soybeans increased farm income levels by \$592 million in 2008. Cumulatively over the period 1997–2008, farm incomes have risen by \$2.74 billion (in nominal terms).

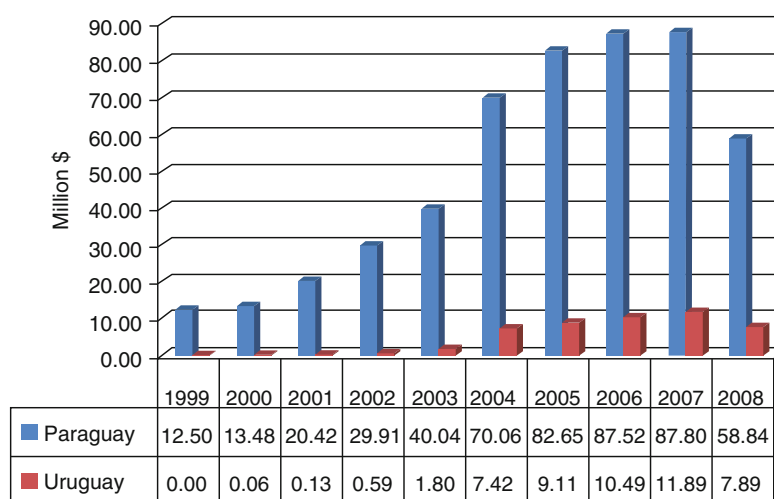
In added value terms, the increase in farm income from the use of the GM HT technology in 2008 was

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 3 Farm-level income impact of using GM HT soybeans in Brazil 1997–2008

Year	Cost savings (\$/ha)	Net cost saving after inclusion of technology cost (\$/ha)	Impact on farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1997	38.8	35.19	3.8	0.06
1998	42.12	38.51	20.5	0.31
1999	38.76	35.15	43.5	0.96
2000	65.32	31.71	43.7	0.85
2001	46.32	42.71	58.7	1.02
2002	40.00	36.39	66.7	1.07
2003	77.00	68.00	214.7	1.62
2004	76.66	61.66	320.9	2.95
2005	73.39	57.23	534.6	5.45
2006	81.09	61.32	730.6	6.32
2007	29.85	8.74	116.3	0.68
2008	64.07	44.44	591.9	2.63

Sources and notes:

1. Impact data based on 2004 comparison data from the Parana Department of Agriculture [11] Cost of production comparison: biotech and conventional soybeans, in USDA GAIN report BR4629 of 11 November 2004. www.fas.usad.gov/gainfiles/200411/146118108.pdf for the period to 2006 [11]. From 2007 based on Galveo [12]
2. Cost of the technology from 2003 is based on the royalty payments officially levied by the technology providers. For years up to 2002, the cost of technology is based on costs of buying new seed in Argentina (the source of the seed). This probably overstates the real cost of the technology and understates the cost savings
3. All values for prices and costs denominated in Brazilian Real have been converted to US dollars at the annual average exchange rate in each year



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 1

National farm income benefit from using GM herbicide-tolerant (GM HT) soybeans in Paraguay and Uruguay 1999–2008 (million \$)

equivalent to an annual increase in production of +2.6% (about 1.54 million tons).

Paraguay and Uruguay

GM HT soybeans have been grown since 1999 and 2000 respectively in Paraguay and Uruguay. In 2008, they accounted for 90% of total soybean plantings in Paraguay and 99% of the soybean plantings in Uruguay (as in Argentina, the majority of plantings are to farm saved or uncertified seed). Using the farm-level impact data obtained from the Argentine research [9, 10] – we are not aware of any published country-specific impact research having been conducted in these two countries) and applying this to production in these two countries, Fig. 1 summarizes the national farm-level income benefits that have been derived from using the technology. In 2008, the respective national farm income gains were \$58.8 million in Paraguay and \$7.9 million in Uruguay.

Canada

GM HT soybeans were first planted in Canada in 1997. In 2008, the share of total plantings accounted for by GM HT soybeans was 73% (0.88 million hectares).

At the farm level, the main impacts of use have been similar to the impacts in the USA. The average farm income benefit has been within a range of \$14–\$40/ha

and the increase in farm income at the national level was \$12.6 million in 2008 (Table 4). The cumulative increase in farm income since 1997 has been \$116 million (in nominal terms). In added value terms, the increase in farm income from the use of the GM HT technology in 2008 was equivalent to an annual increase in production of about 1% (34,500 tons).

South Africa

In 2001, GM HT soybeans were planted commercially in South Africa. In 2008, 184,000 ha (80%) of total soybean plantings were to varieties containing the GM HT trait. In terms of impact at the farm level, net cost savings of between \$5/ha and \$9/ha have been achieved through reduced expenditure on herbicides (Table 5), although in 2008, with the significant increase in glyphosate prices relative to other herbicides, this has fallen back to \$2/ha. At the national level, the increase in farm income was \$0.32 million in 2008. Cumulatively, the farm income gain since 2001 has been \$4.13 million.

Romania

In 2008, Romania was not officially permitted to plant GM HT soybeans, having joined the EU at the start of 2007 (the EU has not permitted the growing of GM HT

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 4 Farm-level income impact of using GM HT soybeans in Canada 1997–2008

Year	Cost savings (\$/ha)	Net cost saving/increase in gross margin (inclusive of technology cost: \$/ha)	Impact on farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1997	64.28	41.17	0.041	0.01
1998	56.62	35.05	1.72	0.3
1999	53.17	31.64	6.35	1.29
2000	53.20	31.65	6.71	1.4
2001	49.83	29.17	9.35	3.4
2002	47.78	27.39	11.92	2.79
2003	49.46	14.64	7.65	1.47
2004	51.61	17.48	11.58	1.48
2005	55.65	18.85	13.30	2.26
2006	59.48	23.53	17.99	2.22
2007	61.99	24.52	16.87	1.57
2008	56.59	14.33	12.61	1.03

Sources and notes:

1. Impact data based on George Morris Centre Report [13] and updated in recent years to reflect changes in herbicide prices
2. All values for prices and costs denominated in Canadian dollars have been converted to US dollars at the annual average exchange rate in each year

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 5 Farm-level income impact of using GM HT soybeans in South Africa 2001–2008

Year	Cost savings (\$/ha)	Net cost saving/increase in gross margin after inclusion of technology cost (\$/ha)	Impact on farm income at a national level (\$ millions)
2001	26.72	7.02	0.042
2002	21.82	5.72	0.097
2003	30.40	7.90	0.24
2004	34.94	9.14	0.46
2005	36.17	9.12	1.42
2006	33.96	5.17	0.83
2007	32.95	5.01	0.72
2008	25.38	1.77	0.32

Sources and notes:

1. Impact data (Data source: Monsanto South Africa – data provision not a reference)
2. All values for prices and costs denominated in South African Rand have been converted to US dollars at the annual average exchange rate in each year

soybeans to date). The impact data presented below therefore covers the period 1999–2006.

The growing of GM HT soybeans in Romania had resulted in substantially greater net farm income gains per hectare than any of the other countries using the technology:

- Yield gains of an average of 31% have been recorded [14]. This yield gain has arisen from the substantial improvements in weed control (weed infestation levels, particularly of difficult to control weeds such as Johnson grass have been very high in Romania. This is largely a legacy of the economic transition during the 1990s, which resulted in very low levels of farm income, abandonment of land, and very low levels of weed control. As a result, the weed bank developed substantially and has been subsequently very difficult to control, until the GM HT soybean system became available [glyphosate has been the key to controlling difficult weeds like Johnson grass]). In recent years, as fields have been cleaned up of problem weeds, the average yield gains have decreased and were reported at +13% in 2006 (source: farmer survey conducted in 2006 on behalf of Monsanto Romania).
- The cost of the technology to farmers in Romania tended to be higher than other countries, with seed being sold in conjunction with the herbicide. For example, in the 2002–2006 period, the average cost of seed and herbicide per hectare was \$120–\$130/ha. This relatively high cost however, did not deter adoption of the technology because of the major yield gains, improvements in the quality of soybeans produced (less weed material in the beans sold to crushers that resulted in price premia being obtained in the early years – no longer relevant post 2005), and cost savings derived.
- The average net increase in gross margin in 2006 was \$59/ha (an average of \$105/ha over the 8 years of commercial use: Table 6).
- At the national level, the increase in farm income amounted to \$7.6 million in 2006. Cumulatively in the period 1999–2006, the increase in farm income was \$44.6 million (in nominal terms).
- The yield gains in 2006 were equivalent to an 9% increase in national production (the annual average increase in production over the 8 years was equal to 10.1%).

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 6 Farm-level income impact of using herbicide-tolerant soybeans in Romania 1999–2006

Year	Cost saving (\$/ha)	Cost savings net of cost of technology (\$/ha)	Net increase in gross margin (\$/ha)	Impact on farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1999	162.08	2.08	105.18	1.63	4.0
2000	140.30	–19.7	89.14	3.21	8.2
2001	147.33	–0.67	107.17	1.93	10.3
2002	167.80	32.8	157.41	5.19	14.6
2003	206.70	76.7	219.01	8.76	12.7
2004	63.33	8.81	135.86	9.51	13.7
2005	64.54	9.10	76.16	6.69	12.2
2006	64.99	9.10	58.79	7.64	9.3

Sources and notes:

1. Impact data (Sources: Brookes [14] and Monsanto Romania [15]. Average yield increase 31% applied to all years to 2003 and reduced to +25% 2004, +19% 2005 and +13% 2006. Average improvement in price premia from high quality 2% applied to years 1999–2004
2. All values for prices and costs denominated in Romanian Lei have been converted to US dollars at the annual average exchange rate in each year
3. Technology cost includes cost of herbicides
4. The technology was not permitted to be planted from 2007 – due to Romania joining the EU

- In added value terms, the combined effect of higher yields, improved quality of beans, and reduced cost of production on farm income in 2006 was equivalent to an annual increase in production of 9.3% (33,230 tons).

Mexico

GM HT soybeans were first planted commercially in Mexico in 1997 (on a trial basis) and in 2008, a continued trial area of 7,330 ha (out of total plantings of 88,000 ha) were varieties containing the GM HT trait.

At the farm level, the main impacts of use have been a combination of yield increase (+9.1% in 2004 and 2005, +3.64% in 2006, +3.2% 2007, and +2.4% 2008) and (herbicide) cost savings. The average farm income benefit has been within a range of \$54–\$89/ha (inclusive of yield gain, cost savings, and after payment of the technology fee/seed premium of \$34.5/ha) and the increase in farm income at the national level was \$0.04 million in 2008 (Table 7). The cumulative increase in farm income since 2004 has been \$3.35 million (in nominal terms). In added value terms, the increase in farm income from the use of the GM HT technology in 2008 was equivalent to an annual increase in production of about 0.5%.

Bolivia

GM HT soybeans were officially permitted to be planted in 2008, although “illegal” plantings have

occurred for several years. For the purposes of analysis in this section, impacts have been calculated back to 2005, when an estimated 0.3 million hectares of soybeans used GM HT technology. In 2008, an estimated 453,000 ha (63% of total crop) used GM HT technology.

The main impacts of the technology are as follows (Table 8):

- An increase in yield arising from improved yield control. The research work conducted by Fernandez et al. [19] estimated a 30% yield difference between GM HT and conventional soybeans although some of the yield gain reflected the use of poor-quality conventional seed by some farmers. In the analysis presented, a more conservative yield gain of +15% has been used.
- GM HT soybeans are assumed to trade at a price discount to conventional soybeans of –2.7%, reflecting the higher price set for conventional soybeans by the Bolivian government in 2008.
- The cost of the technology to farmers has been about \$3.3/ha and the cost savings equal to about \$9.3/ha, resulting in a net cost of production change of +\$6/ha.
- Overall, in 2008, the average farm income gain from using GM HT soybeans was about \$80/ha, resulting in a total farm income gain of \$36.3 million. Cumulatively since 2005, the total farm income gain is estimated at \$83.4 million.

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 7 Farm-level income impact of using GM HT soybeans in Mexico 2004–2008

Year	Cost savings (\$/ha)	Net cost saving/increase in gross margin (inclusive of technology cost and yield gain: \$/ha)	Impact on farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
2004	49.44	82.34	1.18	3.07
2005	51.20	89.41	0.94	2.13
2006	51.20	72.98	0.51	1.05
2007	51.05	66.84	0.33	0.9
2008	33.05	54.13	0.40	0.5

Sources and notes:

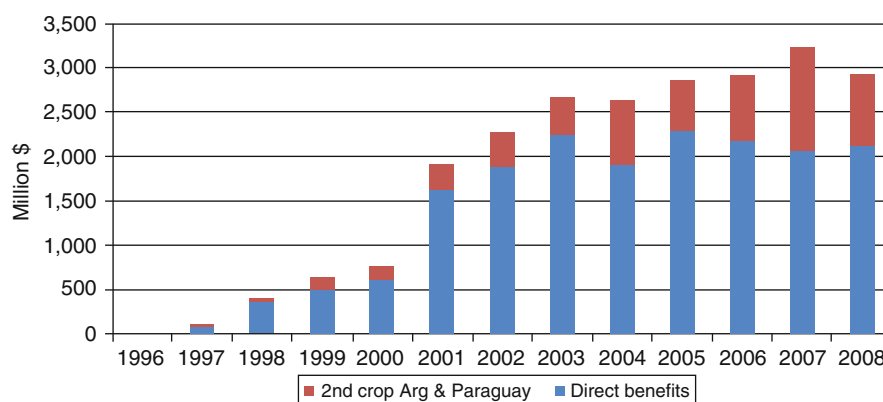
1. Impact data based on Monsanto, 2005, 2007, and 2008 [16–18]. Reportes final del programa Soya Solución Faena en Chiapas. Monsanto Comercial
2. All values for prices and costs denominated in Mexican pesos have been converted to US dollars at the annual average exchange rate in each year

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 8 Farm-level income impact of using GM HT soybeans in Bolivia 2005–2008

Year	Cost savings excluding seed cost premium (\$/ha)	Net cost saving/increase in gross margin (inclusive of technology cost and yield gain: \$/ha)	Impact on farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
2005	9.28	39.73	12.08	4.09
2006	9.28	36.60	15.55	6.35
2007	9.28	44.40	19.45	7.37
2008	9.28	80.09	36.33	7.24

Sources and notes:

1. Impact data based on Fernandez et al. [19]. Average yield gain assumed +15%, cost of technology \$3.32/ha



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 2

Global farm-level income benefits derived from using GM HT soybeans 1996–2008 (million \$)

Summary of Global Economic Impact

In global terms, the farm-level impact of using GM HT technology in soybeans was \$2.12 billion in 2008 (Fig. 2). If the second crop benefits arising in Argentina are included, this impact rises to \$2.92 billion. Cumulatively since 1996, the farm income benefit has been (in nominal terms) \$17.9 billion (\$23.3 billion if second crop gains in Argentina and Paraguay are included).

In terms of the total value of soybean production from the countries growing GM HT soybeans in 2008, the additional farm income (inclusive of Argentine second crop gains) generated by the technology is equal to a value-added equivalent of 4.3%. Relative to the value of global soybean production in 2008, the farm income benefit added the equivalent of 4.1%.

These economic benefits should be placed within the context of a significant increase in the level of soybean production in the main GM adopting countries since 1996 (a 63% increase in the area planted in the leading soybean producing countries of the USA, Brazil, and Argentina).

These economic benefits mostly derive from cost savings although farmers in Mexico, Bolivia, and Romania also obtained yield gains (from significant improvements in weed control levels relative to levels applicable prior to the introduction of the technology). If it is also assumed that all of the second crop soybean gains are effectively additional production that would not have otherwise occurred without the GM HT technology (the GM HT technology facilitated major expansion of second crop soybeans in Argentina and

to a lesser extent in Paraguay) then these gains are de facto “yield” gains. Under this assumption, of the total cumulative farm income gains from using GM HT soy, \$5.56 billion (24%), is due to yield gains/second crop benefits and the balance, 76%, is due to cost savings.

Herbicide-Tolerant Maize

The USA

Herbicide-tolerant maize has been used commercially in the USA since 1997 and in 2008 was planted on 63% of the total US maize crop. The impact of using this technology at the farm level is summarized in Fig. 3. As with herbicide-tolerant soybeans, the main benefit has been to reduce costs, and hence improve profitability levels. Average profitability improved by \$20–\$25/ha in most years (\$17.6/ha in 2008 – affected by the significant increase in glyphosate prices relative to other herbicides). The net gain to farm income in 2008 was \$354 million and cumulatively, since 1997 the farm income benefit has been \$1.7 billion. In added value terms, the effect of reduced costs of production on farm income in 2008 was equivalent to an annual increase in production of 0.71% (2.17 million tons).

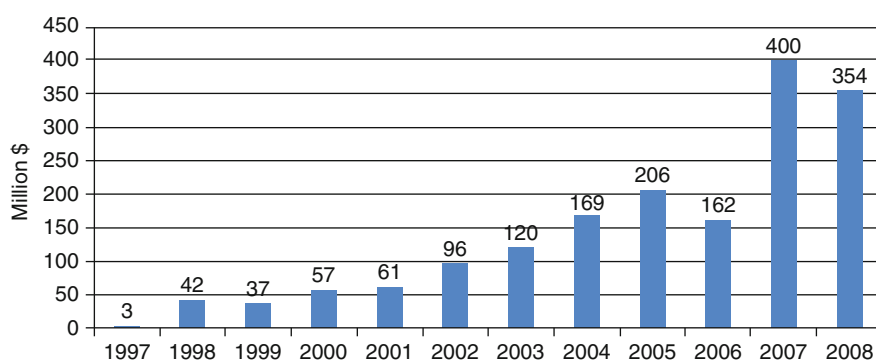
Canada

In Canada, GM HT maize was first planted commercially in 1997. By 2008, the proportion of total plantings

accounted for by varieties containing a GM HT trait was 51%. As in the USA, the main benefit has been to reduce costs and to improve profitability levels. Average annual profitability has improved by between \$12/ha and \$18/ha up to 2007, but fell to about \$6/ha in 2008 (due to the higher price increases for glyphosate relative to other herbicides). In 2008, the net increase in farm income was \$3.7 million and cumulatively since 1999 the farm income benefit has been \$45.8 million. In added value terms, the effect of reduced costs of production on farm income in 2008 was equivalent to an annual increase in production of 0.22% (23,500 tons: Fig. 4).

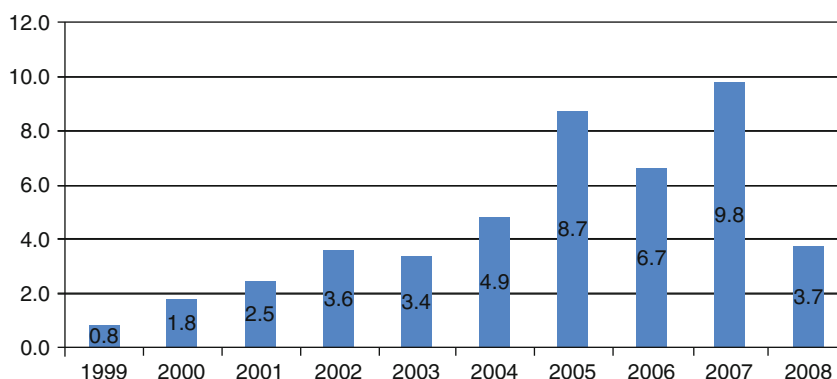
Argentina

GM HT maize was first planted commercially in Argentina in 2004 and in 2008, varieties containing a GM HT trait were planted on 805,000 ha (35% of the total maize area). It has been adopted in two distinct types of area, the majority (80%) in the traditional “corn production belt” and 20% in newer maize-growing regions, which have been traditionally known as more marginal areas that surround the “Corn Belt.” The limited adoption of GM HT technology in Argentina up to 2006 was mainly due to the technology only being available as a single gene, not stacked with the GM IR trait, which most maize growers have also adopted. Hence, faced with an either GM HT or GM IR trait available for use, most farmers have chosen the



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 3

National farm income impact of using GM HT maize in the USA 1997–2008 (Source and notes: Impact analysis based on Sankala and Blumenthal [6, 7] and Johnson and Strom [8] and updated for 2008 to reflect changes in herbicide prices. Estimated cost of the technology \$14.83/ha in years up to 2004, \$17.3/ha in 2005, \$24.71/ha 2006 onward. Cost savings (mostly from lower herbicide use) \$33.47/ha in 2004, \$38.61/ha 2005, \$29.27/ha 2006, \$42.28/ha 2007, and \$40.87/ha 2008)



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 4

National farm income impact of using GM HT maize in Canada 1999–2008 (\$ million) (Source and notes: Impact analysis based on data supplied by Monsanto Canada. Estimated cost of the technology \$18–\$32/ha, cost savings (mostly from lower herbicide use) \$31–\$45/ha)

GM IR trait because the additional returns derived from adoption have tended to be (on average) greater from the GM IR trait than the GM HT trait (see below for further details of returns from the GM HT trait). Stacked traits became available in 2007 and contributed to the significant increase in the GM HT maize area relative to 2006.

In relation to impact on farm income, the following observations were made:

- In all regions, the cost of the technology (about \$20/ha) has been broadly equal to the saving in herbicide costs.
- In the Corn Belt area, use of the technology has resulted in an average 3% yield improvement via improved weed control. In the more marginal areas, the yield impact has been much more significant (+22%) as farmers have been able to significantly improve weed control levels.
- In 2008, the additional farm income at a national level from using GM HT technology has been +\$61.6 million, and cumulatively since 2004, the income gain has been \$113.8 million.

South Africa

Herbicide-tolerant maize has been grown commercially in South Africa since 2003, and 6,46,000 ha out of total plantings of 2.43 million hectares were herbicide tolerant in 2008. Farmers using the technology

have found that small net savings in the cost of production have occurred (i.e., the cost saving from reduced expenditure on herbicides has been greater than the cost of the technology), although in 2008, due to the significant rise in the global price of glyphosate relative to their herbicides, the net farm income balance was negative, at about –\$2/ha. This resulted in a total net farm loss arising from using GM HT technology of \$1.43 million, though since 2003, there has been a net cumulative income gain of \$3.77 million.

Philippines

GM HT maize was first grown commercially in 2006, and 2008 was planted on 270,000 ha. Information about the impact of the technology is limited, although industry sources estimate that, on average farmers using it have derived a 15% increase in yield. Based on a cost of the technology of \$24–\$27/ha (and assuming no net cost savings), the net national impact on farm income was +\$15.9 million in 2008. Cumulatively, since 2006, the total farm income gain has been \$27.1 million

Summary of Global Economic Impact

In global terms, the farm-level economic impact of using GM HT technology in maize was \$433.5 million in 2008 (82% of which was in the USA). Cumulatively since 1997, the farm income benefit has been

(in nominal terms) \$1.9 billion. Of this, 92% has been due to cost savings and 8% to yield gains (from improved weed control relative to the level of weed control achieved by farmers using conventional technology).

In terms of the total value of maize production in the main countries using this technology in 2008, the additional farm income generated by the technology is equal to a value-added equivalent of 0.3% of global maize production.

Herbicide-Tolerant Cotton

The USA

GM HT cotton was first grown commercially in the USA in 1997 and in 2008 was planted on 68% of total cotton plantings.

The farm income impact of using GM HT cotton is summarized in Table 9. The primary benefit has been to reduce costs, and hence improve profitability levels, with annual average profitability increasing by between \$21/ha and \$49/ha (the only published source that has

examined the impact of HT cotton in the USA is the work by Sankala and Blumenthal [6, 7], and Johnson and Strom [8]. In the 2001 study, the costs saved were based on historic patterns of herbicides used on conventional cotton in the mid/late 1990s. The latter studies estimated cost savings on the basis of the conventional herbicide treatment that would be required to deliver the same level of weed control as GM HT cotton. Revised analysis has, however, been conducted for 2008 to reflect changes in the costs of production (notably cost of the technology (in particular “Roundup Ready Flex technology”), higher prices for glyphosate relative to other herbicides in 2008 and additional costs incurred to control weeds resistant to glyphosate in some regions) in the years up to 2004. Since then, net income gains have fallen to between \$1/ha and \$5/ha. The relatively small positive impact on direct farm income in 2008 (and in the last few years) reflects a combination of reasons, including the higher cost of the technology, significant price increases for glyphosate relative to price increases for other

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 9 Farm-level income impact of using GM HT cotton in the USA 1997–2008

Year	Cost savings (\$/ha)	Net cost saving/increase in gross margins, inclusive of cost of technology (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1997	34.12	21.28	12.56	0.2
1998	34.12	21.28	30.21	0.58
1999	34.12	21.28	53.91	1.29
2000	34.12	21.28	61.46	1.22
2001	65.59	45.27	161.46	4.75
2002	65.59	45.27	153.18	3.49
2003	65.59	45.27	129.75	2.33
2004	83.35	48.80	154.72	2.87
2005	71.12	2.89	9.57	0.18
2006	73.66	3.31	13.29	0.22
2007	76.01	5.40	16.56	0.32
2008	72.76	1.20	2.50	0.08

Source and notes:

1. Impact analysis based on Sankala and Blumenthal [6, 7] and Johnson and Strom [8] and own analysis for 2008
2. Estimated cost of the technology \$12.85/ha (1997–2000) and \$21.32/ha 2001–2003, \$34.55 2004, \$68.22/ha 2005, \$70.35/ha 2006, \$70.61/ha 2007, and £71.56/ha 2008

herbicides, and additional costs incurred for management of weeds resistant to glyphosate (notably *Palmer Amaranth*). Overall, the net direct farm income impact in 2008 is estimated to be \$2.5 million (this does not take into consideration any nonpecuniary benefits associated with adoption of the technology: see Section 3.9). Cumulatively, since 1997, there has been a net farm income benefit from using the technology of \$799 million.

Other Countries

Australia, Argentina, South Africa, and Mexico are the other countries where GM HT cotton is commercially grown; from 2000 in Australia, 2001 in South Africa, 2002 in Argentina, and 2005 in Mexico. In 2008, 79% (50,460 ha), 38% (124,000 ha), 75% (9,750 ha), and 40% (50,000 ha) respectively of the total Australian, Argentine, South African, and Mexican cotton crops were planted to GM HT cultivars.

We are not aware of any published research into the impact of GM HT cotton in South Africa, Argentina, or Mexico. In Australia, although research has been conducted into the impact of using GM HT cotton (e.g., Doyle et al. [20]) this does not provide quantification of the impact. Drawing on industry source estimates, the main impacts are as follows:

- *Australia*: No yield gain and cost of the technology in the range of \$30–\$45/ha up to 2007. The cost of the technology increased with the availability of “Roundup Ready Flex” and in 2008 was about \$63/ha. The cost savings from the technology (after taking into consideration the cost of the technology have delivered small net gains of \$5–\$7/ha, although estimates relating to the net average benefits from Roundup Ready Flex are about \$25/ha in 2008 [20]. Overall, in 2008, the total farm income from using the technology was about \$3 million and cumulatively, since 2000, the total gains have been \$8.3 million.
- *Argentina*: No yield gain and a cost of technology in the range of \$30–\$40/ha, although with the increasing availability of stacked traits in recent years, the “cost” part of the HT technology has fallen to \$24/ha. Net farm income gains (after deduction of the cost of the technology) have been \$8–\$18/ha and in 2008 were just under \$10/ha. Overall, in

2008, the total farm income from using GM HT cotton technology was about \$7.4 million, and cumulatively since 2002, the farm income gain has been \$34.2 million.

- *South Africa*: No yield gain and a cost of technology in the range of \$15–\$25/ha. Net farm income gains from cost savings (after deduction of the cost of the technology) have been \$30–\$60/ha. In 2008, the average net gain was \$33.6/ha and the total farm income benefit of the technology was \$0.37 million. Cumulatively since 2001, the total farm income gain from GM HT cotton has been \$2.2 million.
- *Mexico*: Average yield gains of +3.6% from improved weed control have been reported in the first 3 years of use, although no yield gain was recorded in 2008. The average cost of the technology has been in the range of \$60–\$66/ha and typical net farm income gains of about \$80/ha, though in 2008, with no yield gains this fell back to \$16/ha. Overall, in 2008, the total farm income gain from using GM HT cotton was about \$1.35 million and cumulatively since 2005, the total farm income gain has been \$11.7 million.

Summary of Global Economic Impact

Across the five countries using GM HT cotton in 2008, the total farm income impact derived from using GM HT cotton was +\$14.6 million. Cumulatively since 1997, there have been net farm income gains of \$855.8 million (93% of this benefit has been in the USA). Of this, 96% has been due to cost savings and 4% to yield gains (from improved weed control relative to the level of weed control achieved using conventional technology).

Herbicide-Tolerant Canola

Canada

Canada was the first country to commercially use GM HT canola in 1996. Since then, the area planted to varieties containing GM HT traits has increased significantly, and in 2008 was 83% of the total crop (5.43 million hectares).

The farm-level impact of using GM HT canola in Canada since 1996 is summarized in Table 10. The key features are as follows:

- The primary impact in the early years of adoption was increased yields of almost 11% (e.g., in 2002

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 10 Farm-level income impact of using GM HT canola in Canada 1996–2008

Year	Cost savings (\$/ha)	Cost savings inclusive of cost of technology (\$/ha)	Net cost saving/increase in gross margins (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1996	28.59	−4.13	45.11	6.23	0.4
1997	28.08	−4.05	37.11	21.69	1.17
1998	26.21	−3.78	36.93	70.18	3.43
1999	26.32	−3.79	30.63	90.33	5.09
2000	26.32	−3.79	22.42	59.91	5.08
2001	25.15	−1.62	23.10	53.34	5.69
2002	24.84	−3.59	29.63	61.86	6.17
2003	28.04	−4.05	41.42	132.08	6.69
2004	21.42	+4.44	19.09	70.72	4.48
2005	23.11	+4.50	32.90	148.12	6.56
2006	34.02	+16.93	50.71	233.13	8.09
2007	35.44	+17.46	66.39	341.44	7.54
2008	36.36	+17.56	66.63	364.23	6.35

Sources and notes:

1. Impact data based on Canola Council study [21] to 2003 and Gusta et al. [22]. Includes a 10.7% yield improvement and a 1.27% increase in the price premium earned (cleaner crop with lower levels of weed impurities) until 2003. After 2004, the yield gain has been based on differences between average annual variety trial results for Clearfield and biotech alternatives. The biotech alternatives have also been differentiated into glyphosate tolerant and glufosinate tolerant. This resulted in the following observation: for GM glyphosate-tolerant varieties no yield difference for 2004, 2005, and 2008 and +4% 2006 and 2007. For GM glufosinate-tolerant varieties, the yield differences were +12% 2004 and 2008, +19% 2005, +10% 2006 and 2007

2. Negative values denote a net increase in the cost of production (i.e., the cost of the technology was greater than the other cost (e.g., on herbicides) reductions

3. All values for prices and costs denominated in Canadian dollars have been converted to US dollars at the annual average exchange rate in each year

this yield increase was equivalent to an increase in total Canadian canola production of nearly 7%). In addition, a small additional price premium was achieved from crushers through supplying cleaner crops (lower levels of weed impurities). With the development of hybrid varieties using conventional technology, the yield advantage of GM HT canola relative to conventional alternatives (the main one of which is “Clearfield” conventionally derived herbicide-tolerant varieties. Also, hybrid canolas now account for the majority of plantings (including some GM hybrids) with the hybrid vigor delivered by conventional breeding techniques (even in the GM HT [to glyphosate] varieties) has been eroded.

As a result, our analysis has applied the yield advantage of +10.7% associated with the GM HT technology in its early years of adoption (source: Canola Council study of 2001) to 2003. From 2004, the yield gain has been based on differences between average annual variety trial results for “Clearfield” (conventional herbicide-tolerant varieties) and biotech alternatives. The biotech alternatives have also been differentiated into glyphosate tolerant and glufosinate tolerant. This resulted in the following observation: for GM glyphosate-tolerant varieties no yield difference for 2004, 2005, and 2008 and +4% 2006 and 2007. For GM glufosinate-tolerant varieties, the yield differences

were +12% 2004 and 2008, +19% 2005, +10% 2006 and 2007. The quality premia associated with cleaner crops (see above) has not been included in the analysis from 2004.

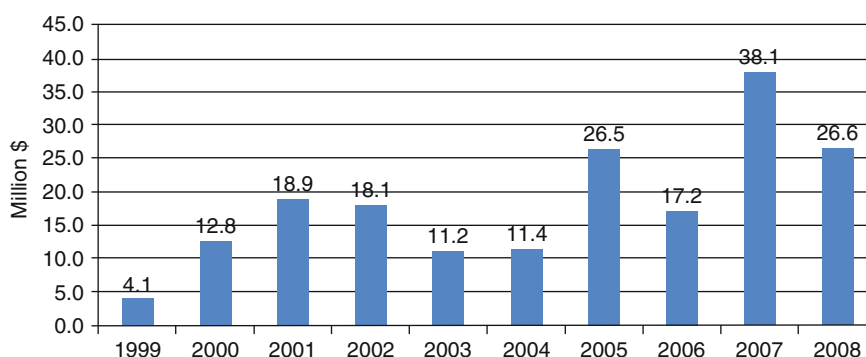
- Cost of production (excluding the cost of the technology) has fallen, mainly through reduced expenditure on herbicides and some savings in fuel and labor. These savings have annually been between about \$25/ha and \$36/ha. The cost of the technology to 2003 was however marginally higher than these savings resulting in a net increase in costs of \$3–\$5/ha. On the basis of comparing GM HT canola with “Clearfield” HT canola (from 2004), there has been a net cost saving of between \$5/ha and \$10/ha, although in 2008 this was \$17/ha.
- The overall impact on profitability (inclusive of yield improvements and higher quality) has been an increase of between \$22/ha and \$48/ha up to 2003. On the basis of comparing GM HT canola with “Clearfield” HT canola (from 2004), the net increase in profitability has been between \$23/ha and \$66/ha.
- The annual total national farm income benefit from using the technology has risen from \$6 million in 1996 to \$364 million in 2008. The cumulative farm income benefit over the 1996–2008 period (in nominal terms) was \$1.64 billion.
- In added value terms, the increase in farm income in 2008 has been equivalent to an annual increase in production of 6.3%.

The USA

GM HT canola has been planted on a commercial basis in the USA since 1999. In 2008, 95% of the US canola crop was GM HT (380,230 ha).

The farm-level impact has been similar to the impact identified in Canada. More specifically, the following observations were noted:

- Average yields increased by about 6% in the initial years of adoption. As in Canada (see above) the availability of high-yielding hybrid conventional varieties has eroded some of this yield gain in recent years relative to conventional alternatives. As a result, the positive yield impacts post 2004 have been applied on the same basis as in Canada (comparison with Clearfields: see [Canada](#) above).
- The cost of the technology has been \$12–\$17/ha for glufosinate-tolerant varieties and \$12–\$33/ha for glyphosate-tolerant varieties. Cost savings (before inclusion of the technology costs) have been \$35–\$45/ha (\$22/ha in 2008) for glufosinate-tolerant canola and \$40–\$79/ha for glyphosate-tolerant canola.
- The net impact on gross margins has been between +\$22/ha and +\$90/ha (\$5/ha in 2008) for glufosinate-tolerant canola, and +\$28/ha and +\$61/ha for glyphosate-tolerant canola.
- At the national level, the total farm income benefit in 2008 was \$26.6 million ([Fig. 5](#)) and the cumulative benefit since 1999 has been \$185 million.



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 5

National farm income impact of using GM HT canola in the USA 1999–2008 (Source and notes: Impact analysis based on Sankala and Blumenthal [6, 7] and Johnson and Strom [8]. Decrease in total farm income impact 2002–2004 is due to decline in total plantings of canola in the USA (from 612,000 in 2002 to 316,000 ha in 2004). Positive yield impact applied in the same way as Canada from 2004)

- In added value terms, the increase in farm income in 2008 has been equivalent to an annual increase in production of about 10.3%.

Australia

GM HT canola was permitted for commercial use in the two states of Victoria and New South Wales in 2008, and was planted on 10,100 ha in that year (2008/09). Ninety-five percent of these plantings had tolerance to the herbicide glyphosate and the balance were tolerant to glufosinate.

A fairly comprehensive farm survey-based analysis of impact of the glyphosate-tolerant canola was commissioned by Monsanto, which involved interviews with 92 of the 108 farmers using this technology in 2008/09 [23, 24]. Key findings from this survey are as follows:

- The technology was made available in both open-pollinated and hybrid varieties, with the open-pollinated varieties representing the cheaper end of the seed market, where competition was mainly with open-pollinated varieties containing herbicide tolerance (derived conventionally) to herbicides in the triazine (TT) group. The hybrid varieties containing glyphosate tolerance competed with nonherbicide-tolerant conventional hybrid varieties and herbicide-tolerant “Clearfield” hybrids (tolerant to the imidazolinone group of herbicides), although, were used in 2008, all of the 33 farmers in the survey using GM HT hybrids did so mainly in competition and comparison with “Clearfield” varieties.
- The GM HT open-pollinated varieties sold to farmers at a premium of about \$Aus3/ha (about \$2.5 US/ha) relative to the TT varieties. The GM HT hybrids sold at a seed premium of about \$Aus 9/ha (\$7.55 US/ha) compared to “Clearfield” hybrids. In addition, farmers using the GM HT technology paid a “technology” fee in two parts; one part was a set fee of \$Aus500 per farm plus \$Aus 10.2/ton of output of canola. On the basis that there were 108 farmers using GM HT (glyphosate tolerant) technology in 2008, the average “up front” fee paid for the technology was \$Aus5.62/ha. On the basis of average yields obtained for the two main types of GM HT seed used, those using open-

pollinated varieties paid \$11.83/ha (basis: average yield of 1.16 tons/ha) and those using GM HT hybrids paid \$Aus12.95/ha (basis: average yield of 1.27 tons/ha). Therefore, the total seed premium and technology fee paid by farmers for the GM HT technology in 2008–2009 was \$Aus20.45/ha (\$17.16 US/ha) for open-pollinated varieties and \$Aus 27.57/ha (\$23.13 US/ha) for hybrid varieties. After taking into consideration, the seed premium/technology fees, the GM HT system was marginally more expensive by \$Aus3/ha (\$2.5 US/ha) and \$Aus4/ha (\$US 3.36/ha) respectively for weed control than the TT and Clearfield varieties.

- The GM HT varieties delivered higher average yields than their conventional counterparts: +22.11% compared to the TT varieties and +4.96% compared to the “Clearfield” varieties. In addition, the GM HT varieties produced higher oil contents of +2% and +1.8% respectively compared to TT and “Clearfield” varieties.
- The average reduction in weed control costs from using the GM HT system (excluding seed premium/technology fee) was \$Aus 17/ha for open-pollinated varieties (competing with TT varieties) and \$Aus 24/ha for hybrids (competing with Clearfield varieties).

In the analysis summarized below in Table 11, these research findings have been applied to the total GM HT

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 11 Farm-level income impact of using GM HT canola in Australia 2008 (\$US)

Year	Average cost saving (\$/ha)	Average cost savings (net after cost of technology) (\$/ha)	Average net increase in gross margins (\$/ha)	Increase in farm income at a national level (\$)
2008	19.18	−20.77	93.37	943,054

Source derived from and based on Monsanto survey of license holders 2008

Notes:

1. The average values shown are weighted averages
2. Other weighted average values derive include yield +21.1% and quality (price) premium of 2.1% applied on the basis of this level of increase in average oil content

crop area on a weighted basis in which the results of GM HT open-pollinated varieties that compete with TT varieties were applied to 64% of the total area and the balance of area used the results from the GM HT hybrids competing with “Clearfield” varieties. This weighting reflects the distribution of farms in the survey, in which 59 (64%) of the farmers indicated they grew open-pollinated varieties and 33 (34%) grew hybrids. The findings show an average farm income gain of \$US 93/ha and a total farm income gain of \$0.93 million in 2008.

Summary of Global Economic Impact

In global terms, the farm-level impact of using GM HT technology in canola in Canada, the USA, and Australia was \$392 million in 2008. Cumulatively, since 1996, the farm income benefit has been (in nominal terms) \$1.83 billion. Within this, 79% has been due to yield gains and the balance (21%) has been from cost savings.

In terms of the total value of canola production in these three countries in 2008, the additional farm income generated by the technology is equal to a value-added equivalent of 6.9%. Relative to the value of global canola production in 2008, the farm income benefit added the equivalent of 1.5%.

GM Herbicide-Tolerant (GM HT) Sugar Beet

GM HT sugar beet was first grown commercially in the USA in 2007 (under 1,000 ha), although it was 2008 before sufficient quantities of seed were available for widespread commercial cultivation. In 2008, just under 258,000 ha of GM HT sugar beet were planted, equal to about 63.5% of the total US crop. The highest levels of penetration of the technology (85% plus of total crop) occurred in Idaho, Wyoming, Nebraska, and Colorado, with about 50% of the crops in the largest sugar beet growing states of North Dakota and Michigan being GM HT.

Impact of the technology in these early years of adoption has been identified as follows:

(a) *Yield*: Analysis by Kniss [25] covering a limited number of farms in Wyoming (2007) identified positive yield impacts of +8.8% in terms of additional root yield (from better weed control) and +12.6% in terms of sugar content relative to conventional crops (i.e., the GM HT crop had about

a 3.8% higher sugar content, which amounts to a 12.8% total sucrose gain relative to conventional sugar beet once the root yield gain was taken into consideration). In contrast, Khan [26] found similar yields reported between conventional and GM HT sugar beet in the Red River Valley region (North Dakota) and Michigan. These contrasting results probably reflect a combination of factors including:

- The sugar beet growing regions in Wyoming can probably be classified as high weed problem areas, and as such, are regions where obtaining effective weed control is difficult using conventional technology (timing of application is key to weed control in sugar beet, with optimal time for application being when weeds are small). Also some weeds (e.g., Kochia) are resistant to some of the commonly used ALS inhibitor herbicides like chlorsulfuron. The availability of GM HT sugar beet with its greater flexibility on application timing has therefore potentially delivered important yield gains for such growers.
 - The GM HT trait was not available in all leading varieties suitable in all growing regions in 2008, hence the yield benefits referred to above from better weed control have to some extent been counterbalanced by only being available in poorer performing germ plasm in states like Michigan and North Dakota (notably not being available in 2008 in leading varieties with rhizomania resistance). It should be noted that the authors of the research cited in this section both perceive that yield benefits from using GM HT sugar beet will be a common feature of the technology in most regions once the technology is available in leading varieties.
 - The year 2008 was reported to have been, in the leading sugar beet growing states, a reasonable year for controlling weeds through conventional technology (i.e., it was possible to get good levels of weed control through timely applications), hence the similar performance reported between the two systems.
- (b) *Costs of production*
- Kniss's work in Wyoming identified weed control costs (comprising herbicides, application,

cultivation, and hand labor) for conventional beet of \$437/ha compared to \$84/ha for the GM HT system. After taking into consideration the \$131/ha seed premium/technology fee for the GM HT trait, the net cost differences between the two systems was \$222/ha in favor of the GM HT system. Kniss did, however, acknowledge that the conventional costs associated with this sample were high relative to most producers (reflecting application of maximum dose rates for herbicides and use of hand labor), with a more typical range of conventional weed control costs being between \$171/ha and \$319/ha (average \$245/ha).

- Khan's analysis puts the typical weed control costs in the Red River region of North Dakota to be about \$227/ha for conventional compared to \$91/ha for GM HT sugar beet. After taking into consideration the seed premium/technology fee (assumed by Khan to be \$158/ha), the total weed control costs were \$249/ha for the GM HT system, \$22/ha higher than the conventional system. Despite this net increase in average costs of production, most growers in this region used (and planned to continue using), the GM HT system because

of the convenience and weed control flexibility benefits associated with it (which research by Marra and Piggot [1]) estimated in the corn, soybean, and cotton sectors to be valued at between \$12/ha and \$25/ha to US farmers). It is also likely that Khan's analysis may understate the total cost savings from using the technology by not taking into account savings on application costs and labor for hand weeding.

For the purposes of our analysis, we have drawn on both these pieces of work, as summarized in Table 12. This shows a net farm income gain in 2008 of over \$21 million to US sugar beet farmers (average gain per hectare of just under \$83/ha). With the availability of GM HT technology in more of the leading varieties, it is expected that the farm income gains associated with yield gains will be greater in subsequent years.

GM Insect-Resistant (To Corn-Boring Pests: GM IR) Maize

The USA

GM IR maize was first planted in the USA in 1996 and in 2008, seed containing GM IR traits was planted on 57% (18.14 million hectares) of the total US maize crop.

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 12 Farm-level income impact of using GM HT sugar beet in the USA 2007–2008

Year	Average cost saving (\$/ha)	Average cost savings (net after cost of technology) (\$/ha)	Average net increase in gross margins (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
2007	353.35	222.39	584.00	472,680	0.03
2008	142.50	−8.58	82.88	21,380,290	1.83

Sources derived from and based on Kniss [25] and Khan [26]

Notes:

1. The yield gains identified by Kniss have been applied to the 2007 GM HT plantings in total and to the estimated GM HT plantings in the states of Idaho, Wyoming, Nebraska, and Colorado, where penetration of plantings in 2008 was 85% (these states account for 26% of the total GM HT crop in 2008), and which are perceived to be regions of above average weed problems. For all other regions, no yield gain is assumed. Across the entire GM HT area in 2008, this equates to a net average yield gain of +3.28%
2. The seed premium of \$131/ha, average costs of weed control respectively for conventional and GM HT systems of \$245/ha and \$84/ha, from Kniss were applied to the crop in Idaho, Wyoming, Nebraska, and Colorado. The seed premium of \$158/ha, weed control costs of \$227/ha and \$249/ha respectively for conventional and GM HT sugar beet, identified by Khan were applied to all other regions using the technology. These states account for 26% of the total GM HT crop in 2008. The resulting average values for seed premium/cost of technology across the entire 2008 GM HT crop was therefore \$151.08/ha and the average weed control cost saving associated with the GM HT system (before taking into consideration the seed premium) was \$142.5/ha

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 13 Farm-level income impact of using GM IR maize in the USA 1996–2008

Year	Cost saving (\$/ha)	Cost savings (net after cost of technology) (\$/ha)	Net increase in gross margins (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1996	24.71	−9.21	29.20	8.76	0.03
1997	24.71	−9.21	28.81	70.47	0.27
1998	20.30	−4.8	27.04	167.58	0.77
1999	20.30	−4.8	25.51	206.94	1.04
2000	22.24	−6.74	24.32	148.77	0.71
2001	22.24	−6.74	26.76	155.87	0.72
2002	22.24	−6.74	30.74	240.45	0.96
2003	22.24	−6.74	31.54	291.00	1.14
2004	15.88	−6.36	33.82	363.41	1.32
2005	15.88	−1.42	34.52	399.91	1.60
2006	15.88	−1.42	55.78	707.23	1.86
2007	15.88	−1.42	61.22	1,136.21	2.28
2008	24.71	−8.83	67.51	1,224.59	2.40

Sources and notes:

1. Impact data based on a combination of studies including the ISAAA (James) review [27], Marra et al. [3], Sankala and Blumenthal [6, 7], and Johnson and Strom [8], Gianessi and Carpenter [28]
2. Yield impact +5% based on average of findings of above studies
3. Insecticide cost savings based on the above references
4. − (minus) value for net cost savings means the cost of the technology is greater than the other cost savings

The farm-level impact of using GM IR maize in the USA since 1996 is summarized in Table 13:

- The primary impact has been increased average yields of about 5% (in 2008 this additional production is equal to an increase in total US maize production of +2.41%).
- The net impact on cost of production has been a small increase of between \$1/ha and \$9/ha (additional cost of the technology being higher than the estimated average insecticide cost savings of \$15–\$16/ha).
- The annual total national farm income benefit from using the technology has risen from \$8.76 million in 1996 to \$1.22 billion in 2008. The cumulative farm income benefit over the 1996–2008 period (in nominal terms) was \$5.12 billion.
- In added value terms, the increase in farm income in 2008 was equivalent to an annual increase in production of 2.4%.

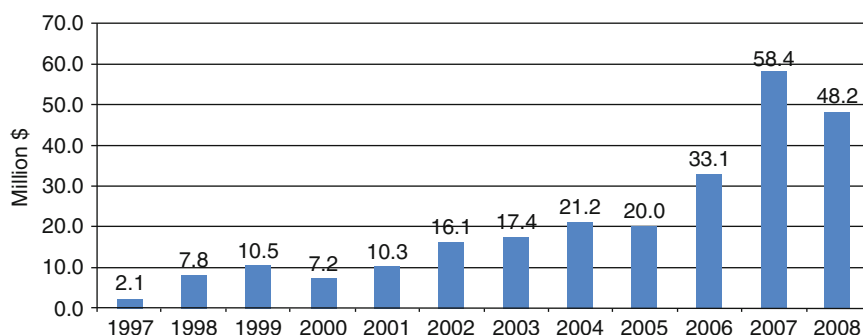
Canada

GM IR maize has also been grown commercially in Canada since 1996. In 2008, it accounted for 62% of the total Canadian maize crop of 1.2 million hectares. The impact of GM IR maize in Canada has been very similar to the impact in the USA (similar yield and cost of production impacts). At the national level, in 2008 the additional farm income generated from the use of GM IR maize was \$48.2 million and cumulatively since 1996 the additional farm income (in nominal terms) was \$252 million (Fig. 6).

Argentina

In 2008, GM IR maize traits were planted on 75% of the total Argentine maize crop (GM IR varieties were first planted in 1998).

The main impact of using the technology on farm profitability has been via yield increases. Various



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 6

National farm income impact of using GM IR maize in Canada 1996–2008

Notes:

1. Yield increase of 5% based on industry assessments (consistent with US analysis). Cost of technology and insecticide cost savings based on US analysis,
2. GM IR area planted in 1996 = 1,000 ha,
3. All values for prices and costs denominated in Canadian dollars have been converted to US dollars at the annual average exchange rate in each year

studies (e.g., see ISAAA review in James [27]) and Trigo and Cap [29] have identified an average yield increase in the region of 8–10%, hence an average of 9% has been used in the analysis up to 2004. More recent trade source estimates provided to the authors put the average yield increased in the last 2–3 years to be between 5% and 6%. Accordingly, our analysis uses a yield increase value of 5.5% for the years from 2004.

No savings in costs of production have arisen for most farmers because very few maize growers in Argentina have traditionally used insecticides as a method of control for corn-boring pests. As such, average costs of production have increased by \$20–\$22/ha (the cost of the technology).

The net impact on farm profit margins (inclusive of the yield gain) has, in recent years, been an increase of about \$20/ha. In 2008, the national level impact on profitability was an increase of \$41 million (an added value equal to 2.15% of the total value of production). Cumulatively, the farm income gain since 1997 has been \$269.7 million.

South Africa

GM IR maize has been grown commercially in South Africa since 2000. In 2008, 56% of the country's total maize crop of 2.42 million hectares used GM IR cultivars.

The impact on farm profitability is summarized in Table 14. The main impact has been an average yield improvement of between 5% and 32% in the years 2000–2004, with an average of about 15% (used as the basis for analysis 2005–2007). In 2008, the estimated yield impact was +10.6% (source: Van der Weld [33]). The cost of the technology \$8–\$17/ha has broadly been equal to the average cost savings from no longer applying insecticides to control corn-boring pests.

At the national level, the increase in farm income in 2008 was \$117.7 million and cumulatively since 2000 it has been \$476 million. In terms of national maize production, the use of GM IR technology on 56% of the planted area has resulted in a net increase in national maize production of 5.9% in 2008. The value of the additional income generated was also equivalent to an annual increase in production of about 5.1%.

Spain

Spain has been commercially growing GM IR maize since 1998 and in 2008, 22% (79,270 ha) of the country's maize crop was planted to varieties containing a GM IR trait.

As in the other countries planting GM IR maize, the main impact on farm profitability has been increased yields (an average increase in yield of 6.3% across farms using the technology in the early years of adoption).

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 14 Farm-level income impact of using GM IR maize in South Africa 2000–2008

Year	Cost savings (\$/ha)	Net cost savings inclusive of cost of technology (\$/ha)	Net increase in gross margin (\$/ha)	Impact on farm income at a national level (\$ millions)
2000	13.98	1.87	43.77	3.31
2001	11.27	1.51	34.60	4.46
2002	8.37	0.6	113.98	19.35
2003	12.82	0.4	63.72	14.66
2004	14.73	0.46	20.76	8.43
2005	15.25	0.47	48.66	19.03
2006	14.32	–2.36	63.75	63.05
2007	13.90	0.22	182.90	225.70
2008	11.74	–4.55	87.07	117.73

Sources and notes:

1. Impact data (Sources: Gouse [30–32] and Van der Weld [33])
2. Negative value for the net cost savings = a net increase in costs (i.e., the extra cost of the technology was greater than the other (e.g., less expenditure on insecticides) cost savings
3. All values for prices and costs denominated in South African Rand have been converted to US dollars at the annual average exchange rate in each year

With the availability and widespread adoption of the Mon 810 trait from 2003, the reported average positive yield impact is about +10%. There has also been a net annual average saving on cost of production (from lower insecticide use) of between \$37/ha and \$61/ha (Table 15). At the national level, these yield gains and cost savings have resulted in farm income being boosted, in 2008 by \$17.9 million and cumulatively since 1998 the increase in farm income (in nominal terms) has been \$77.9 million.

Relative to national maize production, the yield increases derived from GM IR maize were equivalent to a 2.2% increase in national production (2008). The value of the additional income generated from Bt maize was also equivalent to an annual increase in production of 2.1%.

Other EU countries

A summary of the impact of GM IR technology in other countries of the EU is presented in Table 16. This shows that in 2008, the additional farm income derived from using GM IR technology in these six countries was +\$2.5 million, and cumulatively over the 2005–2008 period, the total income gain was \$11.1 million.

Other Countries

GM IR maize has been grown commercially in the following countries:

- *The Philippines* since 2003. In 2008, 280,000 ha out of total plantings of 2.6 million (7%) were GM IR. Estimates of the impact of using GM IR (Sources: Gonsalves [36], Yorobe [37], and Ramon [38]) show annual average yield increases in the range of 14.3–34%. Taking the midpoint of this range (+24.15%), coupled with a small average annual insecticide cost saving of about \$12–\$13/ha and average cost of the technology of about \$33/ha, the net impact on farm profitability has been between \$37/ha and \$109/ha. In 2008, the national farm income benefit derived from using the technology was \$33.5 million and cumulative farm income gain since 2003 has been \$61.2 million.
- *Uruguay* since 2004, and in 2008, 110,000 ha (73% of the total crop) were GM IR. Using Argentine data as the basis for assessing impact, the cumulative farm income gain over the 3 years has been \$3.9 million.
- *Brazil* starting in 2008, when 1.45 million hectares were planted to varieties containing a GM IR trait.

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 15 Farm-level income impact of using GM IR maize in Spain 1998–2008

Year	Cost savings (\$/ha)	Net cost savings inclusive of cost of technology (\$/ha)	Net increase in gross margin (\$/ha)	Impact on farm income at a national level (\$ millions)
1998	37.40	3.71	95.16	2.14
1999	44.81	12.80	102.20	2.56
2000	38.81	12.94	89.47	2.24
2001	37.63	21.05	95.63	1.10
2002	39.64	22.18	100.65	2.10
2003	47.50	26.58	121.68	3.93
2004	51.45	28.79	111.93	6.52
2005	52.33	8.72	144.74	7.70
2006	52.70	8.78	204.5	10.97
2007	57.30	9.55	274.59	20.63
2008	61.49	10.25	225.36	17.86

Sources and notes:

1. Impact data (based on Brookes [34] and Brookes [35]). Yield impact +6.3% to 2004 and 10% used thereafter (originally Bt 176, latterly Mon 810). Cost of technology based on €18.5/ha to 2004 and €35/ha from 2005
2. All values for prices and costs denominated in Euros have been converted to US dollars at the annual average exchange rate in each year

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 16 Farm-level income impact of using GM IR maize in other EU countries 2005–2008

	Year first planted GM IR maize	Area 2008 (hectares)	Yield impact (%)	Cost of technology 2008 (\$/ha)	Cost savings 2008 (before deduction of cost of technology: \$/ha)	Net increase in gross margin 2008 (\$/ha)	Impact on farm income at a national level 2008 (million \$)
France	2005	Nil	N/p	N/p	N/p	N/p	N/p
Germany	2005	3,173	+4	58.57	73.21	78.64	0.25
Portugal	2005	4,851	+12.5	51.24	0	75.60	0.37
Czech Republic	2005	8,380	+10	51.24	26.35	101.95	0.85
Slovakia	2005	1,930	+12.3	51.24	0	228.31	0.44
Poland	2006	3,000	+12.5	51.24	0	133.08	0.40
Romania	2007	7,146	+7.1	46.85	0	26.59	0.19
Total other EU (excluding Spain)		28,480					2.5

Source and notes:

1. Source: Based on Brookes [35]
2. All values for prices and costs denominated in Euros have been converted to US dollars at the annual average exchange rate in each year
3. N/p – planting not permitted in France in 2008

Based on analysis from Galveo [12], the average yield impact was +4.66%, the cost of the technology was \$21.6/ha, insecticide cost savings were \$42/ha, and the average improvement to farm income equal to \$48.12/ha. Overall, the increase in farm profitability associated with the adoption in 2008 was \$69.8 million;

- *Honduras*. Here farm-level “trials” have been permitted since 2003, and in 2008, an estimated 9,000 ha used GM IR traits. Evidence from Falck Zepeda et al. [39] indicated that the primary impact of the technology has been to increase average yields (in 2008 +24%). As insecticides have not traditionally been used by most farmers, no costs of production savings have arisen, coupled with no additional cost for use of the technology (which has been provided free of charge for the trials). In our analysis, we have, however assumed a cost of the technology of \$30/ha, and based on this, the estimated farm income benefit derived from the technology was \$1.1 million in 2008 and cumulatively since 2003 the income gain has been \$2 million.

Summary of Economic Impact

In global terms, the farm-level impact of using GM IR maize was \$1.56 billion in 2008. Cumulatively since 1996, the benefit has been (in nominal terms) \$6.34 billion. This farm income gain has mostly derived from improved yields (less pest damage) although in some countries farmers have derived a net cost saving associated with reduced expenditure on insecticides.

In terms of the total value of maize production from the countries growing GM IR maize in 2008, the additional farm income generated by the technology is equal to a value-added equivalent of 2.2%. Relative to the value of global maize production in 2008, the farm income benefit added the equivalent of 1.2%.

Insect-Resistant (Bt) Cotton (GM IR)

The USA

GM IR cotton has been grown commercially in the USA since 1996 and by 2008, was used in 63% (1.93 million hectares) of total cotton plantings.

The farm income impact of using GM IR cotton is summarized in Table 17. The primary benefit has been increased yields (by 9–11%), although small net savings in costs of production have also been obtained (reduced expenditure on insecticides being marginally greater than the cost of the technology). Overall, average profitability levels increased by \$53–\$115/ha with Bollgard I cotton (with a single Bt gene) between 1996 and 2002 and by between \$87/ha and \$118/ha in 2003–2008 with Bollgard II (containing two Bt genes and offering a broader spectrum of control). This resulted in a net gain to farm income in 2008 of \$189 million. Cumulatively, since 1996, the farm income benefit has been \$2.44 billion. In added value terms, the effect of the increased yields and reduced costs of production on farm income in 2008 was equivalent to an annual increase in production of 6.3% (165,400 tons).

China

China first planted GM IR cotton in 1997, since when the area planted to GM IR varieties has increased to 64% of the total 5.95 million hectares crop in 2008.

As in the USA, a major farm income impact has been via higher yields of 8–10% on the crops using the technology, although there have also been significant cost savings on insecticides used and the labor previously used to undertake spraying. Overall, annual average costs have fallen by about \$145–\$200/ha and annual average profitability improved by \$123–\$472/ha. In 2008, the net national gain to farm income was \$859 million (Table 18). Cumulatively, since 1997, the farm income benefit has been \$7.6 billion. In added value terms, the effect of the increased yields and reduced costs of production on farm income in 2008 was equivalent to an annual increase in production of 17.1% (1.38 million tons).

Australia

Australia planted 83% of its 2008 cotton crop (total crop of 146,000 ha) to varieties containing GM IR traits (Australia first planted commercial GM IR cotton in 1996).

Unlike the other main countries using GM IR cotton, Australian growers have rarely derived yield gains from using the technology (reflecting the effective use

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 17 Farm-level income impact of using GM IR cotton in the USA 1996–2008

Year	Cost savings (net after cost of technology (\$/ha))	Net increase in gross margins (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1996	4.98	115.32	94.69	1.19
1997	4.98	103.47	87.28	1.30
1998	4.98	88.54	80.62	1.47
1999	4.98	65.47	127.29	2.89
2000	4.98	74.11	162.88	3.10
2001	4.98	53.04	125.22	3.37
2002	4.98	69.47	141.86	3.11
2003	5.78	120.49	239.98	4.27
2004	5.78	107.47	261.23	4.82
2005	24.48	117.81	332.41	5.97
2006	−5.77	86.61	305.17	4.86
2007	−2.71	114.50	296.00	5.49
2008	−2.71	98.22	189.50	5.89

Sources and notes:

1. Impact data based on Gianessi and Carpenter [28], Sankala and Blumenthal [6, 7], Johnson and Strom [8], Marra et al. [3], and Mullins and Hudson [40]
2. Yield impact +9% 1996–2002 Bollgard I and +11% 2003 onward Bollgard II
3. Cost of technology: 1996–2002 Bollgard I \$58.27/ha, 2003–2004 Bollgard II \$68.32/ha, \$49.62/ha 2005, \$46.95/ha 2006, \$25.7/ha 2007 and 2008
4. Insecticide cost savings \$63.26/ha 1996–2002, \$74.10/ha 2003–2005, \$41.18/ha 2006, \$28.4/ha 2007 and 2008

of insecticides for pest control prior to the availability of GM IR cultivars), with the primary farm income benefit being derived from lower costs of production (Table 19). More specifically, the following observations were made:

- In the first 2 years of adoption of the technology (Ingard, single gene Bt cotton), small net income losses were derived, mainly because of the relatively high price charged for the seed. Since this price was lowered in 1998, the net income impact has been positive, with cost saving of between \$54/ha and \$90/ha, mostly derived from lower insecticide costs (including application) more than offsetting the cost of the technology.
- For the last few years of use, Bollgard II cotton (containing 2 Bt genes) has been available offering effective control of a broader range of cotton pests.

Despite the higher costs of this technology, users have continued to make significant net cost savings of \$186–\$212/ha.

- At the national level in 2008, the net farm income gains were \$24.2 million and cumulatively since 1996 the gains have been \$214.9 million.
- In added value terms, the effect of the reduced costs of production on farm income in 2008 was equivalent to an annual increase in production of 37% (105,000 tons).

Argentina

GM IR cotton has been planted in Argentina since 1998. In 2008, it accounted for 73% of total cotton plantings.

The main impact in Argentina has been yield gains of 30% (which has resulted in a net increase in total cotton production (2008) of 22%). This has more than

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 18 Farm-level income impact of using GM IR cotton in China 1997–2008

Year	Cost savings (net after cost of technology (\$/ha)	Net increase in gross margins (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1997	194	333	11.33	0.13
1998	194	310	80.97	1.15
1999	200	278	181.67	4.62
2000	–14	123	150.18	2.61
2001	378	472	1,026.26	20.55
2002	194	327	687.27	11.19
2003	194	328	917.00	12.15
2004	194	299	1,105.26	16.89
2005	145	256	845.58	13.57
2006	146	226	792.28	16.86
2007	152	248	942.7	14.46
2008	148	224	858.6	17.14

Sources and notes:

1. Impact data based on Pray et al. [41, 42], which covered the years 1999–2001. Other years based on average of the 3 years, except 2005 onward based on Shachuan (2006) – personal communication
2. Negative cost savings in 2000 reflect a year of high pest pressure (of pests not the target of GM IR technology), which resulted in above average use of insecticides on GM IR using farms
3. Yield impact +8% 1997–1999 and +10% 2000 onward
4. Negative value for the net cost savings in 2000 = a net increase in costs (i.e., the extra cost of the technology was greater than the savings on insecticide expenditure – a year of lower than average bollworm problems)
5. All values for prices and costs denominated in Chinese Yuan have been converted to US dollars at the annual average exchange rate in each year

offset the cost of using the technology. In terms of gross margin, cotton farmers have gained annually between \$25/ha and \$249/ha during the period 1998–2007. At the national level, the annual farm income gains in the last 5 years have been in the range of \$2–\$27 million (Fig. 7). Cumulatively since 1998, the farm income gain from use of the technology has been \$95.4 million. In added value terms, the effect of the yield increases (partially offset by higher costs of production) on farm income in 2008 was equivalent to an annual increase in production of 14.6%.

Mexico

GM IR cotton has been planted commercially in Mexico since 1996. In 2008, GM IR cotton was planted on 70,000 ha (56% of total cotton plantings).

The main farm income impact of using the technology has been yield improvements of between 6% and 9% over the last 6 years. In addition, there have been important savings in the cost of production (lower insecticide costs). Overall, the annual net increase in farm profitability has been within the range of \$104/ha and \$354/ha between 1996 and 2008 (Table 20). At the national level, the farm income benefit in 2008 was \$10.5 million and the impact on total cotton production was an increase of 5.2%. Cumulatively since 1996, the farm income benefit has been \$76.4 million. In added value terms, the combined effect of the yield increases and lower cost of production on farm income in 2008 was equivalent to an annual increase in production of 5.4%.

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 19 Farm-level income impact of using GM IR cotton in Australia 1996–2008

Year	Cost of technology (\$/ha)	Net increase in gross margins/cost saving after cost of technology (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1996	–191.7	–41.0	–1.63	–0.59
1997	–191.7	–35.0	–2.04	–0.88
1998	–97.4	91.0	9.06	0.43
1999	–83.9	88.1	11.80	4.91
2000	–89.9	64.9	10.71	4.38
2001	–80.9	57.9	7.87	5.74
2002	–90.7	54.3	3.91	3.43
2003	–119.3	256.1	16.3	11.49
2004	–179.5	185.8	45.7	21.33
2005	–229.2	193.4	47.9	23.75
2006	–225.9	190.7	22.49	26.01
2007	–251.33	212.1	11.73	40.90
2008	–264.26	199.86	24.23	37.40

Sources and notes:

1. Impact data based on Doyle [43], Taylor [44], CSIRO [45] for bollgard II since 2004

2. All values for prices and costs denominated in Australian dollars have been converted to US dollars at the annual average exchange rate in each year

South Africa

In 2008, GM IR cotton (first planted commercially in 1998) was planted on 7,750 ha in South Africa (84% of the total crop).

The main impact on farm incomes has been significantly higher yields (an annual average increase of about 24%). In terms of cost of production, the additional cost of the technology (between \$17/ha and \$24/ha for Bollgard I and \$40–\$50/ha for Bollgard II (2006 onward) has been greater than the insecticide cost and labor (for water collection and spraying) savings (\$12–\$23/ha), resulting in an increase in overall cost of production of \$2–\$32/ha. Combining the positive yield effect and the increase in cost of production, the net effect on profitability has been an annual increase of between \$27/ha and \$232/ha.

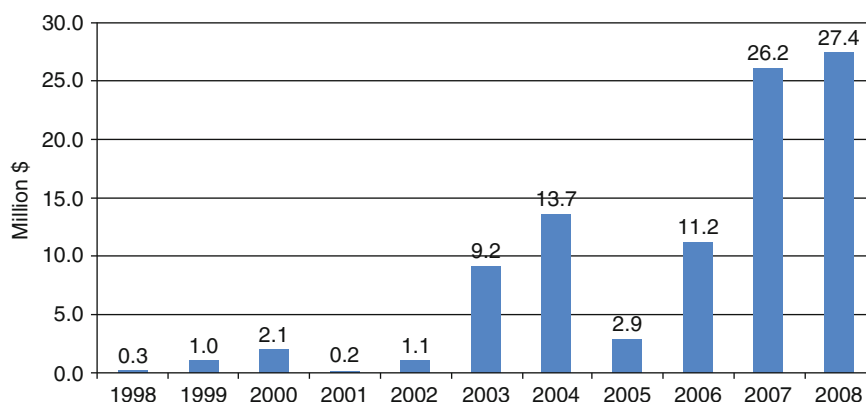
At the national level, farm incomes, over the last 5 years have annually increased by between \$1.2 million and \$1.7 million (Fig. 8). Cumulatively since 1998, the farm income benefit has been \$21 million. The impact

on total cotton production was an increase of 20.1% in 2008. In added value terms, the combined effect of the yield increases and lower costs of production on farm income in 2008 was equivalent to an annual increase in production of 14.5% (based on 2008 production levels).

India

GM IR cotton has been planted commercially in India since 2002. In 2008, 6.97 million hectares were planted to GM IR cotton, which is equal to 77% of total plantings.

The main impact of using GM IR cotton has been major increases in yield [54] found average yield increases of 45% in 2002 and 63% in 2003 (average over the 2 years of 54%) relative to conventionally produced cotton. More recent survey data from Monsanto [16] confirm this high-yield impact (+58% reported in 2004) as do data from IMRB [55], which found an average yield increase of 64% in 2005, and IMRB [56], which found a yield impact of +50% in 2006. With respect to cost of production, the average cost of the



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 7

National farm income impact of using GM IR cotton in Argentina 1998–2008

Sources and notes:

1. Impact data (Sources: Qaim and De Janvry [46, 47]), Elena [48] and for 2005 and 2006 Monsanto LAP, although cost of technology in 2005 from Monsanto Argentina. Area data: source ArgenBio
2. Yield impact +30%, cost of technology \$86/ha (\$40/ha 2005), cost savings (reduced insecticide use) \$17.47/ha
3. All values for prices and costs denominated in Argentine Pesos have been converted to US dollars at the annual average exchange rate in each year

technology (seed premium: \$49–\$54/ha) up to 2006 was greater than the average insecticide cost savings of \$31–\$58/ha resulting in a net increase in costs of production. Following the reduction in the seed premium in 2006 to about \$20/ha, farmers have, on average made a net cost saving of about \$25/ha. Coupled with the yield gains, important net gains to levels of profitability have been achieved of between \$82/ha and \$356/ha. At the national level, the farm income gain in 2008 was \$1.79 billion and cumulatively since 2002 the farm income gains have been \$5.14 billion (Table 21).

The impact on total cotton production was an increase of 31% in 2008 and in added value terms, the combined effect of the yield increases and higher costs of production on farm income in 2008 was equivalent to an annual increase in production of 24% (based on the 2008 production level that is inclusive of the GM IR related yield gains).

Brazil

GM IR cotton was planted commercially in Brazil for the first time in 2006, and in 2008 was planted on 178,000 ha (20% of the total crop). This represents a fall in the share of total plantings relative to 2007, when GM IR traits were planted on 32% of the crop. This decline in

plantings largely reflects the relative performance of the seed containing the GM IR traits compared to the leading conventional varieties, in which the GM IR trait has not been available. In 2006, on the basis of industry estimates of impact of GM IR cotton relative to similar varieties, an average yield gains of +6% and a net cost saving (reduced expenditure on insecticides after deduction of the premium paid for using the technology) of about +\$25/ha were realized. In 2007 and 2008, however, analysis by Galveo [12] and Monsanto Brazil [57] suggests that the yield performance of the varieties containing GM IR traits has been lower (by –3.6% and –2.7% respectively for 2007 and 2008). As a result, the net farm income of using the technology was (after taking into consideration insecticide cost savings and the seed premium), on average, –\$34.5/ha in 2007 and a small net gain of about \$2/ha in 2008. At a national level in 2008, GM IR cotton technology delivered a net gain of about \$0.35 million (a net loss of \$12.3 million in 2007). Cumulatively, the total farm income impact has been positive at about \$5 million.

Other Countries

- *Colombia:* GM IR cotton has been grown commercially in Colombia since 2002 (20,000 ha planted in

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 20 Farm-level income impact of using GM IR cotton in Mexico 1996–2008

Year	Cost savings (net after cost of technology (\$/ha)	Net increase in gross margins (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
1996	58.1	354.5	0.32	0.1
1997	56.1	103.4	1.72	0.5
1998	38.4	316.4	11.27	2.71
1999	46.5	316.8	5.27	2.84
2000	47.0	262.4	6.85	5.76
2001	47.6	120.6	3.04	3.74
2002	46.1	120.8	1.84	3.81
2003	41.0	127.7	3.33	3.67
2004	39.3	130.4	6.24	4.51
2005	40.8	132.3	10.4	7.64
2006	20.4	124.4	6.44	4.06
2007	20.5	139.7	8.38	4.74
2008	19.9	150.4	10.52	5.44

Sources and notes:

1. Impact data based on Traxler et al. [49] covering the years 1997 and 1998. Yield changes data in other years based on official reports submitted to the Mexican Ministry of Agriculture by Monsanto Comercial (Mexico). Also, Martinez-Carillo and Diaz-Lopez [50]
2. Yield impacts: 1996 +37%, 1997 +3%, 1998 +20%, 1999 +27%, 2000 +17%, 2001 +9%, 2002 +7%, 2003 +6%, 2004 +7.6%, 2005 onward +9.25%
3. All values for prices and costs denominated in Mexican Pesos have been converted to US dollars at the annual average exchange rate in each year

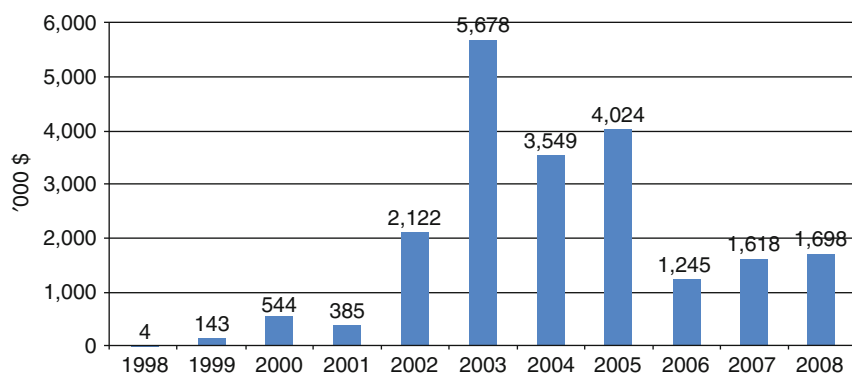
2008 out of a total cotton crop of 40,000 ha). Drawing on recent analysis of impact by Zambrano et al. [58], this shows the main impact has been through a significant improvement in yields of +32%. On the cost impact side, this analysis shows that farmers using GM IR cotton tend to have substantially higher expenditures on pest control than their conventional counterparts, which when taking into consideration the approximate \$70/ha cost of the technology results in a net addition to costs of between \$200/ha and \$280/ha each year (relative to typical expenditures by conventional cotton growers). Nevertheless, after taking into consideration the positive yield effects, the net impact on profitability has been positive. In 2008, the average improvement in profitability was about \$33/ha and the total net gain from using the technology was \$0.91 million.

Cumulatively, since 2002, the net farm income gain has been \$13.9 million.

- *Burkino Faso*: GM IR cotton was grown commercially first in 2008. Based on analysis of pre-commercial trials by Vitale et al. [59, 60], the main impact of the technology is improved yields (by +20%) and savings in insecticide expenditure of about \$62/ha. Based on a cost of technology of about \$42/ha, the net cost savings are about \$20/ha, and inclusive of the yield gains, the estimated net income gain in 2008 was \$124/ha. The total aggregate farm income gain in 2008 was therefore \$1 million.

Summary of Global Impact

In global terms, the farm-level impact of using GM IR cotton was \$2.9 billion in 2008. Cumulatively, since



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 8

National farm income impact of using GM IR cotton in South Africa 1998–2008

Sources and notes:

1. Impact data based on Ismael et al. [51], Kirsten et al. [52], Morse et al. [53]
2. Yield impact +24%, cost of technology \$14–\$24/ha for Bollgard I and about \$50/ha for Bollgard II, cost savings (reduced insecticide use) \$12–\$23/ha
3. All values for prices and costs denominated in South African Rand have been converted to US dollars at the annual average exchange rate in each year
4. The decline in the total farm income benefit post 2003 relative to earlier years reflects the decline in total cotton plantings. This was caused by relatively low farm-level prices for cotton in 2004 and 2005 (reflecting a combination of relatively low world prices and a strong South African currency). In more recent years, cotton has become less competitive relative to alternatives such as corn because of higher world grain prices

1996, the farm income benefit has been (in nominal terms) \$15.61 billion. Within this, 65% of the farm income gain has derived from yield gains (less pest damage) and the balance (35%) from reduced expenditure on crop protection (spraying of insecticides).

In terms of the total value of cotton production from the countries growing GM IR in 2008, the additional farm income generated by the technology is equal to a value-added equivalent of 19.3% (based on the 2008 production level inclusive of the GM IR related yield gains). Relative to the value of global cotton production in 2008, the farm income benefit added the equivalent of 11.1%.

Other Biotech Crops

Maize/Corn Rootworm Resistance

GM rootworm-resistant (CRW) corn has been planted commercially in the USA since 2003. In 2008, there were 13.7 million hectares of CRW corn (43% of the total US crop).

The main farm income impact (Impact data based on Sankala and Blumenthal [6, 7], Johnson and Strom

[8], Rice [61]), and Alston et al. [62]) has been higher yields of about 5% relative to conventional corn. The impact on average costs of production has been +\$2/ha to –\$10/ha (based on an average cost of the technology of \$35–\$42/ha and an insecticide cost saving of \$32–\$37/ha). As a result, the net impact on farm profitability has been +\$28/ha to +\$79/ha.

At the national level, farm incomes increased by \$4.6 million in 2003, rising to \$1.1 billion in 2008. Cumulatively since 2003, the total farm income gain from the use of CRW technology in the USA corn crop has been \$2 billion.

CRW cultivars were also planted commercially for the first time in 2004 in Canada. In 2008, the area planted to CRW-resistant varieties was 119,380 ha. Based on US costs, insecticide cost savings and yield impacts, this has resulted in additional income at the national level of \$8.65 million in 2008 (cumulative total since 2004 of \$13 million).

At the global level, the extra farm income derived from biotech CRW maize use since 2003 has been just over \$2 billion. In 2008, the additional farm income

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 21 Farm-level income impact of using GM IR cotton in India 2002–2008

Year	Cost savings (net after cost of technology (\$/ha))	Net increase in gross margins (\$/ha)	Increase in farm income at a national level (\$ millions)	Increase in national farm income as % of farm-level value of national production
2002	–12.42	82.66	3.69	0.26
2003	–16.2	209.85	20.98	0.47
2004	–13.56	193.36	96.68	1.86
2005	–22.25	255.96	332.74	5.26
2006	3.52	221.02	839.89	14.04
2007	26.41	356.85	2,093.97	22.84
2008	24.28	256.73	1,790.16	24.27

Sources and notes:

1. Impact data based on Bennett et al. [54] and IMRB [55, 56]. As 2008 was reported to be a year of below average pest pressure, the average yield gain used was reduced to +40%
2. All values for prices and costs denominated in Indian Rupees have been converted to US dollars at the annual average exchange rate in each year

generated from use of the technology was equal to 0.9% of the value of the global maize crop.

Virus-Resistant Papaya

Ring spot-resistant papaya has been commercially grown in the USA (State of Hawaii) since 1999, and in 2008 (85% of the state's papaya crop was GM virus resistant (700 ha).

The main farm income impact of this biotech crop has been to significantly increase yields relative to conventional varieties. Compared to the average yield in the last year before the first biotech cultivation (1998), the annual average yield increase of biotech papaya relative to conventional crops has been within a range of +15% to +77% (29% in 2008). At a state level, this is equivalent to a 25% increase in total papaya production in 2008.

In terms of profitability (Impact data based on Sankala and Blumenthal [6, 7] and Johnson and Strom [8]), the net annual impact has been an improvement of between \$3,000/ha and \$29,000/ha, and in 2008 this amounted to a net farm income gain of \$5,790/ha and an aggregate benefit across the state of \$4 million. Cumulatively, the farm income benefit since 1999 has been \$53.4 million.

Virus-resistant papaya is also reported to have been grown in China in 2008, on 4,500 ha. No impact data on this technology has been identified.

Virus-Resistant Squash

Biotech virus-resistant squash has also been grown in some states of the USA since 2004 and is estimated to have been planted on 2,900 ha in 2008 (17% of the total crop in the USA – mostly found in Georgia and Florida).

Based on analysis from Johnson and Strom [8], the primary farm income impact of using biotech virus-resistant squash has been derived from higher yields, which in 2008, added a net gain to users of \$26 million. Cumulatively, the farm income benefit since 2004 has been \$107 million.

Insect-Resistant Potatoes

GM insect-resistant potatoes were also grown commercially in the USA between 1996 and 2000 (planted on 4% of the total US potato crop in 1999 (30,000 ha). This technology was withdrawn in 2001 when the technology provider (Monsanto) withdrew from the market to concentrate on GM trait development in

maize, soybeans, cotton, and canola. This commercial decision was also probably influenced by the decision of some leading potato processors and fast-food outlets to stop using GM potatoes because of perceived concerns about this issue from some of their consumers, even though the GM potato provided the producer and the processor with a lower cost, higher yielding, and more consistent product. It also delivered significant reductions in insecticide use Carpenter and Gianessi (2002).

Indirect (Nonpecuniary) Farm-Level Economic Impacts

Apart from the tangible and quantifiable impacts on farm profitability presented above, there are other important, more intangible (difficult to quantify) impacts of an economic nature.

Many of the studies of the impact of biotech crops have identified the following reasons as being important influences for adoption of the technology:

Herbicide-Tolerant Crops

- Increased management flexibility and convenience that comes from a combination of the ease of use associated with broad-spectrum, post-emergent herbicides like glyphosate and the increased/longer time window for spraying. This not only frees up management time for other farming activities but also allows additional scope for undertaking off-farm, income-earning activities.
- In a conventional crop, post-emergent weed control relies on herbicide applications before the weeds and crop are well established. As a result, the crop may suffer “knock-back” to its growth from the effects of the herbicide. In the GM HT crop, this problem is avoided because the crop is both tolerant to the herbicide and spraying can occur at a later stage when the crop is better able to withstand any possible “knock-back” effects.
- Facilitates the adoption of conservation or no tillage systems. This provides for additional cost savings such as reduced labor and fuel costs associated with plowing, additional moisture retention, and reductions in levels of soil erosion.
- Improved weed control has contributed to reduced harvesting costs – cleaner crops have resulted in

reduced times for harvesting. It has also improved harvest quality and led to higher levels of quality price bonuses in some regions and years (e.g., HT soybeans and HT canola in the early years of adoption respectively in Romania and Canada).

- Elimination of potential damage caused by soil-incorporated residual herbicides in follow-on crops and less need to apply herbicides in a follow-on crop because of the improved levels of weed control.
- A contribution to the general improvement in human safety (as manifest in greater peace of mind about own and worker safety) from reduced exposure to herbicides and a switch to more environmentally benign products.

Insect-Resistant Crops

- Production risk management/insurance purposes – the technology takes away much of the worry of significant pest damage occurring and is, therefore, highly valued. Piloted in 2008 and more widely operational from 2009, US farmers using stacked corn traits (containing insect resistance and herbicide-tolerant traits) are being offered discounts on crop insurance premiums equal to \$7.41/ha.
- A “convenience” benefit derived from having to devote less time to crop walking and/or applying insecticides.
- Savings in energy use – mainly associated with less use of aerial spraying and less tillage.
- Savings in machinery use (for spraying and possibly reduced harvesting times).
- Higher quality of crop. There is a growing body of research evidence relating to the superior quality of GM IR corn relative to conventional and organic corn from the perspective of having lower levels of mycotoxins. Evidence from Europe (as summarized in Brookes [35] has shown a consistent pattern in which GM IR corn exhibits significantly reduced levels of mycotoxins compared to conventional and organic alternatives. In terms of revenue from sales of corn, however, no premia for delivering product with lower levels of mycotoxins have, to date, been reported although where the adoption of the technology has resulted in reduced frequency of

crops failing to meet maximum permissible fumonisin levels in grain maize (e.g., in Spain), this delivers an important economic gain to farmers selling their grain to the food using sector. GM IR corn farmers in the Philippines have also obtained price premia of 10% [37] relative to conventional corn because of better quality, less damage to cobs and lower levels of impurities.

- Improved health and safety for farmers and farm workers (from reduced handling and use of pesticides, especially in developing countries where many apply pesticides with little or no use of protective clothing and equipment).
- Shorter growing season (e.g., for some cotton growers in India), which allows some farmers to plant a second crop (notably maize) in the same season. Also some Indian cotton growers have reported knock on benefits for beekeepers as fewer bees are now lost to insecticide spraying [63].

Some of the economic impact studies have attempted to quantify some of these benefits (e.g., Qaim and Traxler [9] quantified some of these in Argentina (a \$3.65/ha saving (−7.8%) in labor costs and a \$6.82/ha (−28%) saving in machinery/fuel costs associated with the adoption of GM HT soybeans). Where identified, these cost savings have been included in the analysis presented above. Nevertheless, it is important to recognize that these largely intangible benefits are considered by many farmers as a primary reason for adoption of GM technology, and in some cases farmers have been willing to adopt for these reasons alone, even when the measurable impacts on yield and direct costs of production suggest marginal or no direct economic gain.

Since the early 2000s, a number of farmer-survey based studies in the USA have also attempted to better quantify these nonpecuniary benefits. These studies have usually employed contingent valuation techniques to obtain farmers valuations of nonpecuniary benefits. A summary of these findings is shown in (Table 22).

Aggregating the Impact to US Crops 1996–2008

The approach used to estimate the nonpecuniary benefits derived by US farmers from biotech crops over the period 1996–2008 has been to draw on the values identified by Marra and Piggot ([1, 2]; Table 22) and

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 22 Values of nonpecuniary benefits associated with biotech crops in the USA

Survey	Median value (\$/ha)
2002 IR (to rootworm) corn growers survey	7.41
2002 soybean (HT) farmers survey	12.35
2003 HT cropping survey (corn, cotton, and soybeans) – North Carolina	24.71
2006 HT (flex) cotton survey	12.35 (relative to first generation HT cotton)

Source: Marra and Piggot 2006 and 2007 [1, 2]

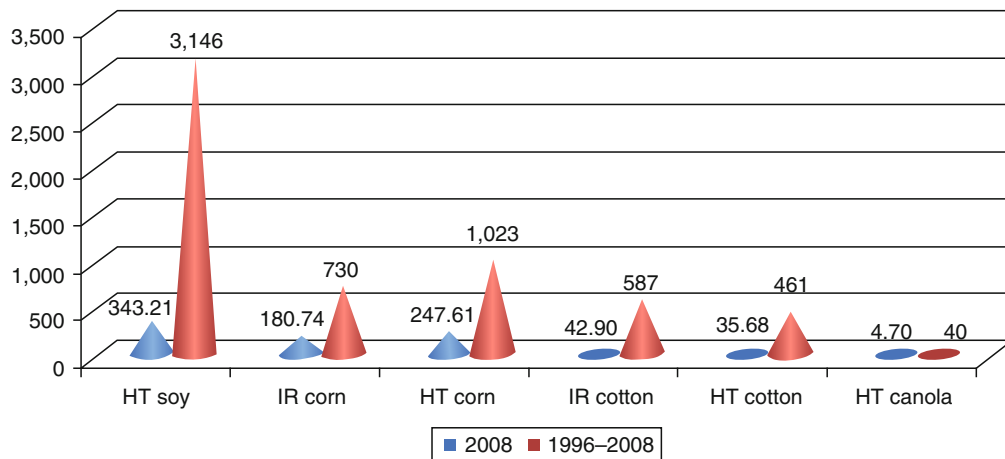
to apply these to the biotech crop planted areas during this 13-year period. Figure 9 summarizes the values for nonpecuniary benefits derived from biotech crops in the USA (1996–2008) and shows an estimated (nominal value) benefit of \$855 million in 2008 and a cumulative total benefit (1996–2008) of \$5.99 billion. Relative to the value of direct farm income benefits presented above, the nonpecuniary benefits were equal to 21% of the total direct income benefits in 2008 and 25.6% of the total cumulative (1996–2008) direct farm income. This highlights the important contribution this category of benefit has had on biotech trait adoption levels in the USA, especially where the direct farm income benefits have been identified to be relatively small (e.g., HT cotton).

Estimating the Impact in Other Countries

It is evident from the literature review that farmers in other countries, who use GM technology, also value the technology for a variety of nonpecuniary/intangible reasons. The most appropriate methodology for identifying these nonpecuniary benefit valuations in other countries would be to repeat the type of US farmer surveys in other countries. Unfortunately, the authors are not aware of any such studies having been undertaken to date.

Production Effects of the Technology

Based on the yield assumptions used in the direct farm income benefit calculations presented above and taking



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 9
Nonpecuniary benefits derived by US farmers 1996–2008 by trait (\$ million)

account of the second soybean crop facilitation in South America, biotech crops have added important volumes to global production of corn, cotton, canola, and soybeans since 1996 (Table 23).

The biotech IR traits, used in the corn and cotton sectors, have accounted for 99% of the additional corn production and almost all of the additional cotton production. Positive yield impacts from the use of this technology have occurred in all user countries (except GM IR cotton in Australia: this reflects the levels of *Heliothis* pest control previously obtained with intensive insecticide use. The main benefit and reason for adoption of this technology in Australia has arisen from significant cost savings (on insecticides) and the associated environmental gains from reduced insecticide use) when compared to average yields derived from crops using conventional technology (such as application of insecticides and seed treatments). Since 1996, the average yield impact across the total area planted to these traits over the 12-year period has been +7.1% for corn traits and +14.8% for cotton traits (Fig. 10).

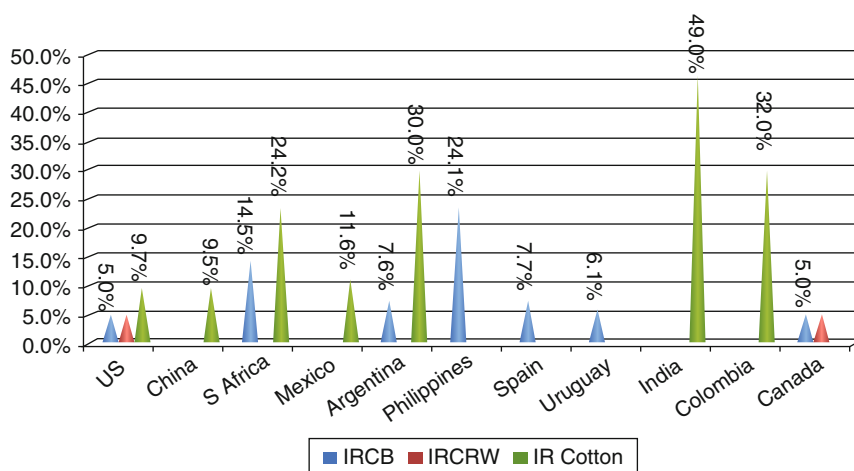
Although the primary impact of biotech HT technology has been to provide more cost-effective (less expensive) and easier weed control versus improving yields from better weed control (relative to weed control obtained from conventional technology), improved weed control has, nevertheless occurred, delivering higher yields in some countries. Specifically, HT soybeans in Romania improved the

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 23 Additional crop production arising from positive yield effects of biotech crops

	1996–2008 additional production (million tons)	2008 additional production (million tons)
Soybeans	74.0	10.1
Corn	79.7	17.1
Cotton	8.6	1.8
Canola	4.8	0.6

average yield by over 30% in early adoption years and biotech HT corn in Argentina and the Philippines delivered yield improvements of +9% and +15% respectively.

Biotech HT soybeans have also facilitated the adoption of no tillage production systems, shortening the production cycle. This advantage enables many farmers in South America to plant a crop of soybeans immediately after a wheat crop in the same growing season. This second crop, additional to traditional soybean production, has added 73.5 million tons to soybean production in Argentina and Paraguay between 1996 and 2008 (accounting for 99% of the total biotech-related additional soybean production).



Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Figure 10

Average yield impact of biotech IR traits 1996–2008 by country and trait Notes: IRCB, resistant to corn-boring pests; IRCRW, resistant to corn rootworm

Using the same sensitivity analysis as applied to the farm income estimates presented in the executive summary to the production impacts (one scenario of consistent lower than average pest/weed pressure and one of consistent higher than average pest/weed pressure), Table 24 shows the range of production impacts.

Summary of Economic Effects of Transgenic/Biotech Crops

Overall, GM technology has had a significant positive impact on farm income derived from a combination of enhanced productivity and efficiency gains (Table 25). In 2008, the direct global farm income benefit from biotech crops was \$9.37 billion. This is equivalent to having added 3.6% to the value of global production of the four main crops of soybeans, maize, canola, and cotton. Since 1996, farm incomes have increased by \$52 billion.

The largest gains in farm income have arisen in the soybean sector, largely from cost savings. The \$2.93 billion additional income generated by GM herbicide-tolerant (GM HT) soybeans in 2008 has been equivalent to adding 4.3% to the value of the crop in the biotech growing countries, or adding the equivalent of 4.1% to the \$71 billion value of the global soybean crop in 2008. These economic benefits should, however be placed within the context of a significant increase in the level of soybean production in the main biotech

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 24 Additional crop production arising from positive yield effects of biotech crops 1996–2008 under different pest/weed pressure assumptions and impacts of the technology (million tons)

Crop	Consistent below average pest/weed pressure	Average pest/weed pressure (main study analysis)	Consistent above average pest/weed pressure
Soybeans	73.8	74.0	74.3
Corn	48.0	79.7	140.9
Cotton	6.2	8.6	11.8
Canola	3.3	4.8	5.2

Note: No significant change to soybean production under all three scenarios as 99% of production gain due to second cropping facilitation of the technology

adopting countries. Since 1996, the soybean area in the leading soybean producing countries of the USA, Brazil, and Argentina increased by 63%.

Substantial gains have also arisen in the cotton sector mainly from the adoption of GM insect-resistant (GM IR) cotton (through a combination of higher yields and lower costs). In 2008, cotton farm income levels in the biotech adopting countries increased by \$2.9 billion and

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 25 Global farm income benefits from growing biotech crops 1996–2008: million US \$

Trait	Increase in farm income 2008	Increase in farm income 1996–2008	Farm income benefit in 2008 as % of total value of production of these crops in biotech adopting countries	Farm income benefit in 2008 as % of total value of global production of crop
GM herbicide-tolerant soybeans	2,925.7	23,342.0	4.3	4.1
GM herbicide-tolerant maize	433.5	1,896.0	0.6	0.3
GM herbicide-tolerant cotton	14.6	855.8	0.1	0.06
GM herbicide-tolerant canola	391.8	1,829.2	6.9	1.5
GM insect-resistant maize	2,645.5	8,344.2	3.7	2.0
GM insect-resistant cotton	2,904.5	15,612.7	19.3	11.1
Others	51.5	162.1	Not applicable	Not applicable
Totals	9,367.1	52,042.0	5.71	3.65

Notes: All values are nominal. Others = Virus-resistant papaya and squash and herbicide-tolerant sugar beet. Totals for the value shares exclude “other crops” (i.e., relate to the four main crops of soybeans, maize, canola, and cotton). Farm income calculations are net farm income changes after inclusion of impacts on yield, crop quality, and key variable costs of production (e.g., payment of seed premia, impact on crop protection expenditure)

since 1996, the sector has benefited from an additional \$15.6 billion. The 2008 income gains are equivalent to adding 19.3% to the value of the cotton crop in these countries, or 11.1% to the \$26 billion value of total global cotton production. This is a substantial increase in value-added terms for two new cottonseed technologies.

Significant increases to farm incomes have also resulted in the maize and canola sectors. The combination of GM insect resistant (GM IR) and GM HT technology in maize has boosted farm incomes by \$10.24 billion since 1996. In the canola sector (largely North American) an additional \$1.83 billion has been generated.

Of the total cumulative farm income benefit, \$31.2 billion (60%) has been due to yield gains (and second crop facilitation), with the balance arising from reductions in the cost of production. Within this yield gain component, 76% derives from the GM IR technology and the balance to GM HT crops.

Table 26 summarizes farm income impacts in key biotech adopting countries. This highlights the important farm income benefit arising from GM HT soybeans in South America (Argentina, Brazil, Paraguay, and Uruguay), GM IR cotton in China and India, and a range of GM cultivars in the USA. It also illustrates

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 26 GM crop farm income benefits 1996–2008 selected countries: million US \$

	GM HT soybeans	GM HT maize	GM HT cotton	GM HT canola	GM IR maize	GM IR cotton	Total
The USA	11,028	1,705.6	799	185.0	7,107	2,444.1	23,268.7
Argentina	8,764.1	113.8	34.2	N/a	269.8	95.4	9,277.3
Brazil	2,745.8	N/a	N/a	N/a	69.8	5.0	2,820.6
Paraguay	503.2	N/a	N/a	N/a	N/a	N/a	503.2
Canada	116.1	45.8	N/a	1,643.2	265.4	N/a	2,070.5
South Africa	4.1	3.8	2.2	N/a	475.8	21.0	506.9
China	N/a	N/a	N/a	N/a	N/a	7,599	7,599
India	N/a	N/a	N/a	N/a	N/a	5,142	5,142
Australia	N/a	N/a	8.3	0.9	N/a	214.9	224.1
Mexico	3.3	N/a	11.7	N/a	N/a	76.1	91.1
The Philippines	N/a	27.1	N/a	N/a	61.2	N/a	88.3
Romania	44.6	N/a	N/a	N/a	N/a	N/a	44.9
Uruguay	49.4	N/a	N/a	N/a	3.9	N/a	53.3
Spain	N/a	N/a	N/a	N/a	77.9	N/a	77.9
Other EU	N/a	N/a	N/a	N/a	11.1	N/a	11.1
Columbia	N/a	N/a	N/a	N/a	N/a	13.9	13.9
Bolivia	83.4	N/a	N/a	N/a	N/a	N/a	83.4

Notes: All values are nominal. Farm income calculations are net farm income changes after inclusion of impacts on yield, crop quality, and key variable costs of production (e.g., payment of seed premia, impact on crop protection expenditure). N/a = not applicable. US total figure excludes \$182.3 million for other crops/traits

the growing level of farm income benefits obtained in South Africa, the Philippines, and Mexico.

In terms of the division of the economic benefits obtained by farmers in developing countries relative to farmers in developed countries, Table 27 shows that in 2008, 50.5% of the farm income benefits have been earned by developing country farmers. The vast majority of these income gains for developing country farmers have been from GM IR cotton and GM HT soybeans. Over the 13 years, 1996–2008, the cumulative farm income gain derived by developing country farmers was also 50% (\$26.2 billion).

Examining the cost farmers pay for accessing GM technology, Table 28 shows that across the four main biotech crops, the total cost in 2008 was equal to 27% of the total technology gains (inclusive of farm income

gains plus cost of the technology payable to the seed supply chain: the cost of the technology accrues to the seed supply chain including sellers of seed to farmers, seed multipliers, plant breeders, distributors, and the GM technology providers).

For farmers in developing countries the total cost was equal to 15% of total technology gains, while for farmers in developed countries the cost was 36% of the total technology gains. While circumstances vary between countries, the higher share of total technology gains accounted for by farm income gains in developing countries relative to the farm income share in developed countries reflects factors such as weaker provision and enforcement of intellectual property rights in developing countries and the higher average level of farm income gain on a per hectare

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 27 GM crop farm income benefits 2008: developing versus developed countries: million US \$

	Developed	Developing
GM HT soybeans	1,232.1	1,693.6
GM IR maize	2,380.5	265.0
GM HT maize	357.4	76.1
GM IR cotton	213.8	2,690.8
GM HT cotton	5.5	9.1
GM HT canola	391.8	0
GM virus-resistant papaya and squash and GM HT sugar beet	51.5	0
Total	4,632.6	4,734.6

Developing countries = all countries in South America, Mexico, Honduras, Burkino Faso, India, China, the Philippines, and South Africa

basis obtained by farmers in developing countries relative to that obtained by farmers in developed countries.

Concluding Comments

Biotechnology has, to date delivered several specific agronomic traits that have overcome a number of production constraints for many farmers. This has resulted in improved productivity and profitability for the 13.3 million adopting farmers who have applied the technology to 115 million hectares in 2008.

During the last 13 years, this technology has made important positive socioeconomic and environmental contributions. These have arisen even though only a limited range of biotech agronomic traits have so far been commercialized, in a small range of crops.

The biotechnology has delivered economic and environmental gains through a combination of their inherent technical advances and the role of the

Global Economic Impact of Transgenic/Biotech Crops (1996–2008). Table 28 Cost of accessing GM technology (million \$) relative to the total farm income benefits 2008

	Cost of technology: all farmers	Farm income gain: all farmers	Total benefit of technology to farmers and seed supply chain	Cost of technology: developing countries	Farm income gain: developing countries	Total benefit of technology to farmers and seed supply chain: developing countries
GM HT soybeans	1,058.2	2,925.7	3,983.9	334.4	1,693.6	2,028.0
GM IR maize	1,045.9	2,645.5	3,691.4	99.7	265.0	364.7
GM HT maize	547.8	433.5	981.3	32.5	76.1	108.6
GM IR cotton	434.6	2,904.5	3,339.1	353.0	2,690.8	3,043.8
GM HT cotton	167.1	14.6	181.7	10.4	9.1	19.5
GM HT canola	109.0	391.8	500.86	N/a	N/a	N/a
Others	41.5	51.5	93.0	N/a	N/a	N/a
Total	3,404.1	9,367.1	12,771.26	830.0	4,734.6	5,564.6

N/a, not applicable. Cost of accessing technology based on the seed premia paid by farmers for using GM technology relative to its conventional equivalents

technology in the facilitation and evolution of more cost-effective and environmentally friendly farming practices. More specifically, it covers the following main issues:

- The gains from the GM IR traits have mostly been delivered directly from the technology (yield improvements, reduced production risk, and decreased the use of insecticides). Thus farmers (mostly in developing countries) have been able to both improve their productivity and economic returns while also practicing more environmentally friendly farming methods.
- The gains from GM HT traits have come from a combination of direct benefits (mostly cost reductions to the farmer) and the facilitation of changes in farming systems. Thus, GM HT technology (especially in soybeans) has played an important role in enabling farmers to capitalize on the availability of a low-cost, broad-spectrum herbicide (glyphosate) and in turn, facilitated the move away from conventional to low/no tillage production systems in both North and South America. This change in production system has made additional positive economic contributions to farmers (and the wider economy) and delivered important environmental benefits, notably reduced levels of GHG emissions (from reduced tractor fuel use and additional soil carbon sequestration).
- Both IR and HT traits have made important contributions to increasing world production levels of soybeans, corn, cotton, and canola.

The impact of GM HT traits has, however contributed to increased reliance on a limited range of herbicides and this has contributed to some limited development of weed resistance to these herbicides. Some degree of reduced effectiveness of glyphosate (and glufosinate) against certain weeds is to be expected and the extent to which this may develop further, will depend on farming practice and behavior relating to mixing, rotation, and sequencing of herbicides. Where resistance has occurred, this has resulted in low-dose rate applications of other herbicides in weed control programs (commonly used in conventional production systems) occurring and hence, has marginally reduced the level of net environmental and economic gains derived from the current use of the

biotechnology. Nevertheless, to date, the overall environmental and economic gains arising from the use of biotech crops have been substantial.

Appendix 1: Argentine Second Crop Soybeans

Year	Second crop area (million hectares)	Increase in income linked to GM HT system (million \$)	Additional production (million tons)
1996	0.45	Negligible	Negligible
1997	0.65	25.4	0.3
1998	0.8	43.8	0.9
1999	1.4	116.6	2.3
2000	1.6	144.2	2.7
2001	2.4	272.8	5.7
2002	2.7	372.6	6.9
2003	2.8	416.1	7.7
2004	3.0	678.1	6.9
2005	2.3	526.7	6.3
2006	3.2	698.9	11.2
2007	4.9	1,133.6	9.88
2008	3.4	764.6	9.62

Additional gross margin based on data from Grupo CEO

Bibliography

1. Marra M, Piggott N (2006) The value of non pecuniary characteristics of crop biotechnologies: a new look at the evidence, North Carolina State University
2. Marra M, Piggott N (2007) The net gains to cotton farmers of a national refuge plan for Bollgard II cotton. *AgBioforum* 10(1):1–10 (www.agbioforum.org)
3. Marra M, Pardey P, Alston J (2002) The pay-offs of agricultural biotechnology: an assessment of the evidence. International Food Policy Research Institute, Washington, USA
4. Carpenter J, Gianessi L (1999) Herbicide tolerant soybeans: why growers are adopting roundup ready varieties. *AgBioforum* 2:65–72
5. Carpenter J (2001) Comparing Roundup ready and conventional soybean yields 1999. National Centre for Food & Agriculture Policy, Washington

6. Sankala S, Blumenthal E (2003) Impacts on US agriculture of biotechnology-derived crops planted in 2003 – an update of eleven case studies. NCFAP, Washington, www.ncfap.org
7. Sankala S, Blumenthal E (2006) Impacts on US agriculture of biotechnology-derived crops planted in 2005 – an update of eleven case studies. NCFAP, Washington, www.ncfap.org
8. Johnson S, Strom S (2008) Quantification of the impacts on US agriculture of biotechnology-derived crops planted in 2006. NCFAP, Washington, www.ncfap.org
9. Qaim M, Traxler G (2002) Roundup ready soybeans in Argentina: farm level, environmental and welfare effects. In: 6th ICABR conference, Ravello, Italy
10. Qaim M, Traxler G (2005) Roundup ready soybeans in Argentina: farm level and aggregate welfare effects. *Agric Econ* 32(1):73–86
11. Parana Department of Agriculture (2004) Cost of production comparison: biotech and conventional soybeans. In: USDA GAIN report BR4629 of 11 November 2004. www.fas.usad.gov/gainfiles/200411/146118108.pdf
12. Galveo A (2009) Farm survey findings of impact of herbicide tolerant soybeans and insect resistant corn and cotton in Brazil, Celeres, Brazil. www.celeres.co.br
13. George Morris Centre (2004) Economic and environmental impacts of the commercial cultivation of glyphosate tolerant soybeans in Ontario (Unpublished report for Monsanto, Canada)
14. Brookes G (2005) The farm level impact of using roundup ready soybeans in Romania. *AgBioforum* 8(4) (Also available on www.pgeconomics.co.uk)
15. Monsanto Romania (2007) Roundup ready soybeans: survey growers crops in 2006 and intentions for 2007
16. Monsanto Comercial Mexico (2005) Official report to Mexican Ministry of Agriculture. (Unpublished)
17. Monsanto Comercial Mexico (2007) Official report to Mexican Ministry of Agriculture of the 2006 crop. (Unpublished)
18. Monsanto Comercial Mexico (2008) Official report to Mexican Ministry of Agriculture of the 2008 cotton crop. (Unpublished)
19. Fernandez W et al (2009) GM soybeans in Bolivia. In: Paper presented to the 13th ICABR conference, Ravello, Italy, June 2009
20. Doyle B et al (2003) The Performance of roundup ready cotton 2001–2002 in the Australian cotton sector. University of New England, Armidale, Australia
21. Canola Council of Canada (2001) An agronomic and economic assessment of transgenic canola. Canola Council, Canada, www.canola-council.org
22. Gusta M et al (2009) Economic benefits of GMHT canola for producers, University of Saskatchewan, College of Biotechnology (Working Paper)
23. Monsanto Australia (2009) Survey of herbicide tolerant canola licence holders 2008
24. Fischer J, Tozer P (2009) Evaluation of the environmental and economic impact of roundup ready canola in the Western Australian crop production system. In: Curtin University of Technology Technical Report 11/2009
25. Kniss A (2009) Farm scale analysis of glyphosate resistant sugar beet in the year of commercial introduction in Wyoming, University of Wyoming
26. Khan M (2008) Roundup ready sugar beet in America. *British Sugar Beet Review* Winter 2008. 76(4):16–19
27. James C (2003) Global review of commercialized transgenic crops 2002: feature Bt maize, ISAAA No 29
28. Gianessi L, Carpenter J (1999) Agricultural biotechnology insect control benefits. NCFAP, Washington, USA
29. Trigo E, Cap E (2006) Ten years of GM crops in Argentine Agriculture, ArgenBio
30. Gouse M et al (2005) A GM subsistence crop in Africa: the case of Bt white maize in S Africa. *Int J Biotechnol* 7(1/2/3)
31. Gouse M et al (2006) Three seasons of insect resistant maize in South Africa: have small farmers benefited. *AgBioforum* 9(1):15–22
32. Gouse M et al (2006) Output and labour effect of GM maize and minimum tillage in a communal area of Kwazulu-Natal. *J Dev Perspect* 2:2
33. Van der Weld W (2009) Final report on the adoption of GM maize in South Africa for the 2008/09 season. South African Maize Trust
34. Brookes G (2003) The farm level impact of using Bt maize in Spain, ICABR conference paper 2003, Ravello, Italy (Also on www.pgeconomics.co.uk)
35. Brookes G (2008) The benefits of adopting GM insect resistant (Bt) maize in the EU: first results from 1998–2006. *Int J Biotechnol* 10(2/3):148–166 (www.pgeconomics.co.uk)
36. Gonsalves D (2005) Harnessing the benefits of biotechnology: the case of Bt corn in the Philippines. ISBN 971-91904-6-9. Strive Foundation, Laguna, Philippines
37. Yorobe J (2004) Economics impact of Bt corn in the Philippines. In: Paper presented to the 45th PAEDA convention, Querson City, Philippines
38. Ramon G (2005) Acceptability survey on the 80–20 bag in a bag insect resistance management strategy Galveo A (2009) Unpublished (in January 2010) data on first survey findings of impact of insect resistant corn (first crop) in Brazil, Celeres, Brazil. www.celeres.co.br
39. Falck Zepeda J et al (2009) Small 'resource poor' countries taking advantage of the new bio-economy and innovation: the case of insect protected and herbicide tolerant corn in Honduras. In: Paper presented to the 13th ICABR conference, Ravello, Italy
40. Mullins W, Hudson J (2004) Bollgard II versus Bollgard sister line economic comparisons. In: 2004 Beltwide cotton conferences, San Antonio, USA, Jan 2004
41. Pray C et al (2001) Impact of Bt cotton in China. *World Dev* 29(5):1–34
42. Pray C et al (2002) Five years of Bt cotton in China – the benefits continue. *Plant J* 31(4): 423–430
43. Doyle B (2005) The performance of Ingard and Bollgard II cotton in Australia during the 2002/2003 and 2003/2004 seasons. University of New England, Armidale, Australia
44. Taylor I (2003) Cotton CRC annual report, UNE, Armidale, Cotton Research Institute, Narrabri, Australia

45. CSIRO (2005) The cotton consultants Australia 2005 Bollgard II comparison report. CSIRO, Australia
46. Qaim M, De Janvry A (2002) Bt cotton in Argentina: analysing adoption and farmers willingness to pay. American Agricultural Economics Association Annual Meeting, California
47. Qaim M, De Janvry A (2005) Bt cotton and pesticide use in Argentina: economic and environmental effects. *Environ Dev Econ* 10:179–200
48. Elena M (2001) Economic advantages of transgenic cotton in Argentina, INTA. As cited in Trigo and Cap (2006)
49. Traxler G et al (2001) Transgenic cotton in Mexico: economic and environmental impacts. In: ICABR conference, Ravello, Italy
50. Martinez-Carillo J, Diaz-Lopez N (2005) Nine years of transgenic cotton in Mexico: adoption and resistance management. In: *Proceedings Beltwide Cotton Conference*, Memphis, USA, June 2005
51. Ismael Y et al (2002) A case study of smallholder farmers in the Mahathini flats, South Africa. In: ICABR conference, Ravello Italy 2002
52. Kirsten J et al (2002) Bt cotton in South Africa: adoption and the impact on farm incomes amongst small-scale and large-scale farmers. In: ICABR conference, Ravello, Italy 2002
53. Morse S et al (2004) Why Bt cotton pays for small-scale producers in South Africa. *Nat Biotechnol* 22(4):379–380
54. Bennett R, Ismael Y, Kambhampati U, Morse S (2004) Economic impact of genetically modified cotton in India. *AgBioforum* 7(3):96–100
55. IMRB (2006) Socio-economic benefits of Bollgard and product satisfaction (in India). IMRB International, Mumbai, India
56. IMRB (2007) Socio-economic benefits of Bollgard and product satisfaction (in India). IMRB International, Mumbai, India
57. Monsanto Brazil (2008) Farm survey of conventional and Bt cotton growers in Brazil 2007. (Unpublished)
58. Zambrano P et al (2009) Insect resistant cotton in Columbia: impact on farmers. In: Paper presented to the 13th ICABR conference, Ravello, Italy, June 2009
59. Vitale J et al (2006) The Bollgard II field trials in Burkina Faso: measuring how Bt cotton benefits West African farmers. In: Paper presented at the 10th ICABR Conference, Ravello, Italy
60. Vitale J et al (2008) The economic impact of 2nd generation Bt cotton in West Africa: empirical evidence from Burkino Faso. *Int J Biotechnol* 10(2/3):167–183
61. Rice M (2004) Transgenic rootworm corn: assessing potential agronomic, economic and environmental benefits. *Plant Health Progress* 10, '094/php-2001-0301-01-RV
62. Alston J et al (2003) An ex-ante analysis of the benefits from adoption of corn rootworm resistant, transgenic corn technology. *AgBioforum* 5(3), Article 1
63. Manjunath T (2008) Bt cotton in India: remarkable adoption and benefits, Foundation for Biotech Awareness and Education, India. www.fbae.org

Global Wind Power Installations

THOMAS ACKERMANN, RENA KUWAHATA
Energynautics GmbH, Langen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Current Status
Types of Wind Turbines
Penetration
Costs
Market Forecast (Future Directions)
Conclusions
Bibliography

Glossary

Clean development mechanism Clean development mechanism is the flexible mechanism under Article 12 of the Kyoto Protocol with the purpose to (1) assist non-Annex I Parties in achieving sustainable development, (2) contribute to the ultimate objective of the UNFCCC (United Nations Framework Convention on Climate Change), and (3) assist Parties included in Annex I achieve compliance with their quantified emission limitation and reduction commitments. Annex I Parties refer to industrialized countries that were members of the OECD (Organization for Economic Co-operation and Development) in 1992 plus countries with economies in transition (the EIT Parties), including the Russian Federation, the Baltic States, and several Central and Eastern European States.

Installed capacity Installed capacity is the total MW of operational generation plant of a given technology.

Offshore Wind power plant installed in a marine environment.

Offshore wind developments Offshore wind developments are wind power plants installed in shallow waters off the coast.

Onshore wind developments Onshore wind developments are wind power plants installed on land.

Definition of the Subject

Wind power has been utilized for over 3,000 years to aid with human activities. Although it was predominantly used as mechanical power at the beginning, gradually, with industrialization, the focus shifted to the use as electrical power. With the arrival of the oil price shock in the early 1970s, the use of wind energy as electrical power started to gain more focus, and since the end of the twentieth-century, wind energy has become one of the most important sustainable energy resources. The installation of wind power production capacity is growing at a rapid pace, doubling every 3 years, and is expected to continue growing in the future, providing around 20% of the world's power needs by 2030. With strong growth in on- and offshore developments in Europe, and rapid market expansions in the rest of the world, it is imperative that the knowledge which is currently concentrated in a few countries be spread, as implications on research, education, and electric utilities all around the globe are significant. This is the main purpose of this chapter, to present an overview of the relevant areas as well as providing links to further readings and related organizations.

Introduction

The power of the wind has been utilized for at least 3,000 years. Until the early twentieth-century, wind power was used to provide mechanical power for pumping water or for grinding grain. At the beginning of modern industrialization, however, the use of fluctuating energy sources such as wind was gradually replaced by more stable power sources such as fossil-fuel-fired engines and electricity.

With the first oil price shock in the early 1970s, interest in the power of the wind resurfaced. This time, however, the main focus was on wind power providing electrical energy rather than mechanical energy. By converting wind energy into electrical energy, it became possible to provide a reliable and consistent power source, particularly when used with other energy technologies as a backup, via the electrical grid.

The first wind turbines for electricity generation had already been developed at the beginning of the twentieth-century; however, technology started improving step by step since the early 1970s. By the end of the 1990s, wind energy reemerged as one of the

most important sustainable energy resources. During the last decade of the twentieth-century, worldwide wind power capacity doubled approximately every 3 years, and in the past decade between 2000 and 2010, it has sustained a compound annual growth rate of over 20% [1]. The cost of wind-generated electricity has fallen by a factor of about 4 during the last 25 years, and some even claim that wind power prices have now converged with the prices of electricity from gas, coal and nuclear plants [2]. Experts in the field predict that the cumulative capacity will be growing worldwide starting off at 27% in 2010 and gradually declining to 5–9% by 2020, supplying up to around 20% of the world's power needs by 2030 [3].

Wind energy technology itself has also moved also very fast toward new dimensions. As can be seen from Table 1, at the end of 1989, a 300-kW wind turbine with a 30-m rotor diameter was state-of-the-art. But only 10 years later, 1,500-kW turbines with a rotor diameter of around 70 m were available from many manufacturers. The first demonstration projects using 2-MW wind turbines with a rotor diameter of 74 m were installed before the turn of the century. 2–5-MW turbines are now in 2010 commercially available, and prototypes of 6–7.5-MW wind turbines are currently being installed. The largest turbine under development is a 15-MW turbine planned for offshore use in 2020 [6].

Global Wind Power Installations. Table 1 Development of wind turbine size between 1985 and 2010

Year	Capacity (kW)	Rotor diameter (m)
1985	50	15
1989	300	30
1992	500	37
1994	600	46
1998	1,500	70
2002	3,000	90
2006	5,000	112
2010	7,000	126
2015	10,000–15,000	150

Source: DEWI [4] and different editions of [5]. 2015 values estimated based on various research proposals

This fast development of the wind energy market as well as of the technology has large implications on research, education, and on professionals working for electric utilities or the wind energy industry. It is important to mention that more than 73% of the worldwide wind capacity is installed in only five countries: USA, China, Germany, Spain, and India. Hence, most of the wind energy knowledge is concentrated in these countries. The use of wind energy technology, however, is fast spreading to other areas in the world. Hence, the available information must also be spread around the world, and this is the main purpose of this chapter. However, despite the fact that wind energy has been already utilized for 3,000 years, it remains a very complex technology. The technology involves a combination of technical disciplines, including aerodynamics, structure-dynamics, mechanical, as well as electrical engineering. Due to the complexity of the wind energy technology, it is not possible to cover all related topics in this chapter in great detail. The chapter aims rather at presenting an overview of the relevant areas as well as providing links to further readings and related organizations.

Current Status

This section will provide a brief overview of the status of wind energy around the world at the end of the first decade of this century. Furthermore, major wind energy support schemes will be presented.

Wind energy statistics are regularly published by various organizations. Regional or countrywide statistics are often compiled and published by the corresponding wind energy associations. For example, The Spanish Wind Energy Association [7], the German Wind Energy Institute [4], the International Economic Platform for Renewable Energies [8], and the Global Wind Energy Council [9] regularly publish worldwide statistics. Up-to date worldwide statistics are published by Windpower Monthly [5] in the January, April, July, and October editions. The Danish wind energy consultant BTM also publishes an annual wind energy development status report with worldwide statistics and forecasts [10]. The latest World Market Update report published in 2010 provides a very good overview of the current status as well as an interesting future scenario of how to meet 10% of the world's electricity demand

with wind power by 2020. The World Wind Energy Association's World Wind Energy Report [11] and the IEA Wind Energy Annual Report [12] provide detailed overviews of the research and industry activities and policies in the area of wind energy for various countries.

Global Status

Wind energy has been the fastest growing energy technology since the 1990s in terms of the percentage of yearly growth of installed capacity per technology source. The growth of wind energy, however, is not evenly distributed around the world (see Table 2). By the end of 2010, around 44% of the worldwide wind energy capacity was installed in Europe, a further 31% in Asia and the Pacific, and 23% in North America.

In 2009, more than 38 GW of new wind power capacity was installed around the world (Table 3). This was a surprise for many people in the industry, who expected the financial crisis to put a halt to the record-breaking trend of the growth in wind energy. In fact, the growth rate was 31.7%, the highest rate since 2001. Over a third of the new installations were made

Global Wind Power Installations. Table 2 Operational wind power capacity worldwide

Region	Installed capacity [MW]				
	End-1995	End-2000	End-2005	End-2009	End-2010
Europe	2,518	12,972	40,898	76,152	86,075
North America	1,676	2,695	9,832	38,383	44,189
Latin America	11	103	212	1,274	2,008
Asia and Pacific	626	1,795	7,878	41,831	61,038
Middle East and Africa	13	141	271	865	1,079
Total world wide	4,844	17,706	59,091	158,505	194,390

Source: 1995 and 2001 January editions of [5], 2006 and 2009 editions of [9], and [13]

Global Wind Power Installations. Table 3 Top-ten new installed capacity 2009

Country	MW	%
China	13,803	36.0
USA	9,996	26.1
Spain	2,459	6.4
Germany	1,917	5.0
India	1,271	3.3
Italy	1,114	2.9
France	1,088	2.8
UK	1,077	2.8
Canada	950	2.5
Portugal	673	1.8
Rest of world	3,994	10.4
Total top 10	34,348	89.6
Total world wide	38,342	100.0

Source: 2009 edition of [9]

by China, more than doubling its installations compared to the previous year, and catapulting itself to second place in total cumulative installed capacity, just ahead of Germany (Table 4). Almost another third was installed in the USA, who maintains its number one position in total cumulative installations. With over 80 countries using wind energy on a commercial basis in the world, wind energy has become an important player in the world's energy markets. Source: 2009 edition of [11].

The top-ten ranking of wind power plant owners and manufacturers in 2009 is shown in Table 5. Leaders in the development of wind power plants are dominated by European companies such as Vestas, Enercon, and Gamesa, which cover 46.7% of total cumulative installed capacity. However, with rapid developments in China, companies such as Sinovel and Goldwind are now breaking into top ranks of the manufacturing industry.

May 26 (Reuters) – China's wind turbine manufacturers are growing fast, challenging the dominance of players such as Vestas and GE.

According to Danish consultants BTM, three Chinese suppliers now rank among the world's top-ten turbine makers.

Global Wind Power Installations. Table 4 Top-ten cumulative installed capacity 2009

Country	MW	%
USA	35,064	22.1
China	25,805	16.3
Germany	25,777	16.3
Spain	19,149	12.1
India	10,926	6.9
Italy	4,850	3.1
France	4,492	2.8
UK	4,051	2.6
Portugal	3,535	2.2
Denmark	3,465	2.2
Rest of world	21,391	13.5
Total top 10	137,114	86.5
Total world wide	158,505	100.0

Source: 2009 edition of [9]

Following is a list of the top-ten wind turbine makers and their percentage of the global market.

Europe

Between the end of 2005 and the end of 2010, around 33% of all new grid-connected wind turbines worldwide were installed in Europe (see Table 2). A breakdown of installed capacity per country is shown in Table 6. The main driver in Europe for this development was the creation of fixed feed-in tariffs in countries such as Germany and Spain. Feed-in tariffs are defined by the government as a price per kWh that distribution companies have to pay for electricity that is generated from renewable resources and fed into the local distribution grid. For an overview of tariffs, see [14]. Fixed feed-in tariffs reduce the risk of fluctuating electricity prices and therefore provide a long-term secure income to investors.

In Denmark, a market price and premium is used for onshore wind power plants, and bidding processes are used for offshore wind power plants. For onshore wind power plants, this means that a feed-in premium is paid for a certain number of hours of electricity produced by wind turbines at the installed output, and in addition, a refund of a smaller amount is paid

Global Wind Power Installations. Table 5 Breakdown of global wind turbine manufacturing industry

No.	Company	Country	Newly installed in 2009 (MW)	%	Cumulative (MW)	%
1	Vestas	Denmark	4,766	12.9%	39,705	23.6%
2	GE wind energy	USA	4,741	12.8%	22,931	13.6%
3	Sinovel	China	3,510	9.5%	5,658	3.4%
4	Enercon	Germany	3,221	8.7%	19,738	11.7%
5	Goldwind	China	2,727	7.4%	5,315	3.2%
6	Gamesa	Spain	2,546	6.9%	19,225	11.4%
7	Dongfang electric	China	2,475	6.7%	3,765	2.2%
8	Suzlon	India	2,421	6.5%	9,671	5.7%
9	Siemens wind power	Denmark/Germany	2,265	6.1%	11,213	6.7%
10	REpower	Germany	1,297	3.5%	4,894	2.9%
Total for other companies			7,034	19.0%	26,331	15.6%
Total			37,003	100.0%	168,446	100.0%
Top-ten companies			29,969	81.0%	142,115	84.4%

Source: GWEC, Wind Power (March 2010) and BTM Consult [5, 9, 10]

for balancing costs of electricity. For offshore wind power plants, potential developers are invited to submit offers for building new wind park projects. Developers bid for the amount of financial incentives to be paid for each kWh fed into the grid by renewable energy systems, and the best bidder(s) is awarded their bid feed-in tariff for a predefined period [15].

In the United Kingdom, the approach is based on Fixed Quotas Combined with Green Certificate Trading also known as ROCs (Renewable Obligation Certificates). This scheme enforces fixed quotas for utilities regarding the amount of renewable energy per year they must sell via their transmission or distribution network. At the same time, producers of renewable energy receive certificates for the amount of renewable energy fed into the grid. Utilities must buy these certificates to show that they have fulfilled their obligation of meeting the quota. Similar schemes are used in a number of other European countries, such as Sweden and Norway [1].

No detailed data regarding the average size of the wind turbines installed in Europe is available. However, for single countries such as Germany, data is available as presented in Table 7. This presents the development of the average size of new wind turbine installations in Germany.

The average size of the yearly installed wind turbines in Germany has increased from 20 kW in 1985 to 1,985-kW in 2010. In 2010, in Germany, 14% of the total newly installed wind turbines had a capacity under 2-MW, whereas the share of turbines with 2–2.9-MW was 82.8%. A small 3% of the newly erected wind turbines had a capacity exceeding 3-MW. On a European scale, turbines of 2-MW and larger have become virtually standard since 2005 [1].

Several countries now have operational offshore wind power plants in Europe. These include Denmark, Sweden, the UK, the Netherlands, Belgium, Ireland, and Finland (see Table 8). Although significant development only started in the early twenty-first century, the growth has been steady, and it is starting to have an increasing impact on Europe's wind power development. In 2010, the annual installed capacity more than doubled compared to the previous year, taking the total installed capacity from 2,063 MW at the end of 2009 to 2,964 MW at the end of 2010 [9, 20].

There are significant plans backed by the European Commission to increase production capacity in the North Sea and develop an offshore grid, which is expected to reach up to 40 GW of installed capacity by 2020 (2009 edition of [9]).

Global Wind Power Installations. Table 6 Operational wind power capacity in Europe

Country	Installed capacity (MW)	
	End-2005	End-2010
Germany	18,428	27,214
Denmark	3,128	3,752
Spain	10,028	20,676
Netherlands	1,224	2,237
UK	1,353	5,204
Sweden	509	2,163
Italy	1,718	5,797
Greece	573	1,208
Ireland	495	1,428
Portugal	1,022	3,702
Austria	819	1,011
France	757	5,660
Belgium	167	911
Turkey	20	1,329
Poland	73	1,107
Other ^a	550	2,677
Total	40,898	86,075

Source: 2006 January edition of [9] and [13]

^aBulgaria, Croatia, Cyprus, Czech Republic, Estonia, Faroe Islands, Finland, Hungary, Iceland, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Norway, Romania, Russia, Slovakia, Slovenia, Switzerland, and Ukraine

Further offshore projects are planned particularly in the UK (Greater Gabbard: 504 MW; Sheringham Shoal: 315 MW; Walney Phase 1: 184 MW; Ormonde: 150 MW), Germany (Bard 1: 400 MW; Baltic 1: 48 MW), Belgium (Bligh Bank: 165 MW) and Italy (Tricase: 90 MW) [16].

A summary of the offshore wind power plants in Europe in 2010 is shown in Table 8.

North America

After the wind power boom in California during the mid-1980s, development of wind energy slowed down significantly in North America. In the middle of the 1990s, the dismantling of old wind power plants

Global Wind Power Installations. Table 7 Development of the average turbine size in Germany

Year	Average size of yearly new installed capacity in Germany (kW)
1985	20
1986	45
1987	44.8
1988	66.9
1989	143.4
1990	164.3
1991	168.8
1992	178.6
1993	255.8
1994	370.6
1995	472.2
1996	530.5
1997	628.9
1998	785.6
1999	935.5
2000	1,114
2001	1,278
2002	1,394
2003	1,553
2004	1,696
2005	1,723
2006	1,849
2007	1,888
2008	1,923
2009	2,013
1 HJ 2010	1,985

Source: Various editions of [4] by DEWI

sometimes even exceeded the installations of new wind turbines, which led to a reduction in the overall installed capacity.

In 1998, a second boom started in the USA. This time, wind project developers aimed at installing projects before the federal Production Tax Credit (PTC) expired on the 30 of June 1999. The PTC

Global Wind Power Installations. Table 8 Offshore wind energy projects

Name	Year	Capacity (MW)	Hub height (m)	Distance from shore (km)	Water depth (m)
Nogersund, SE	1991–1998	1*0.22	37.5	0.25	7
Vindeby, Baltic Sea, DK	1991	11*0.45	37.5	1.5	3–5
Lely, IJsselmeer, NL	1994	4*0.5	41.5	1	5–10
Tunø Knob, Baltic Sea, DK	1995	10*0.5	43	6	3–5
Dronsten, NL	1996	28*600	50	30	1–2
Bockstigen, Baltic Sea, SE	1997	5*0.55	41.5	4	5–6
Utgrunden, Baltic Sea, SE	2000	7*1.425	65	8	7–10
Blyth, North Sea, UK	2000	2*2	58	1	5–6
Middelgrunden Baltic Sea, DK	2001	20*2	60	1–3	2–6
Yttre Stengrund, Baltic Sea, SE	2001	5*2	60	5	8
Horns Rev I, DK	2002	80*2	70	14	6–14
Nysted (Rødsand I)DK	2003	72*2.3	90	6–10	6–0
North Hoyle, UK	2003	30*2	67	3–10	5–12
Samsø, DK	2003	10*2.3		3.5	11–18
Frederikshavn, DK	2003	4*	80	0.8	3
Scroby Sands, UK	2004	30*2	60	2.5	2–10
Arklow Bank, IR	2004	7*3.6	74	10	2.5–5
Enova offshore – Emden, DE	2004	1*4.5	100	<1	–
Kentish Flats, UK	2005	30*3	70	8.5	5
Egmond aan Zee, NL	2006	36*3	70	8–12	19–22
Barrow, UK	2006	30*3	75	7	21–23
Breitling, DE	2006	1*2.5		1	2
Lillgrund, SE	2007	48*2.3	68	10	2.5–9
Beatrice, UK	2007	2*5		25	40
Burbo Bank, UK	2007	25*3.6	84	5.2	10
Lynn and Inner Dowsing, UK	2008	54*3.6	80	5–5.2	10
Princess Amalia, NL	2008	60*2	100	23	19–24
Kemi Ajos I + II, F	2008	10*3	88	<1	3
Hooksiel, DE	2008	1*5	90	0.4	2–8
Thornton Bank I, BE	2008	6*5	94	27–30	12–27
Horns Rev II, DK	2009	90*2.3	68	30	9–17
Rhyl Flats, UK	2009	25*3.6	80	8	4–15
Alpha Ventus, DE	2009	12*5	90	43	30
Storebaelt/Sporgø, DK	2009	7*3		2	6–16

Global Wind Power Installations. Table 8 (Continued)

Name	Year	Capacity (MW)	Hub height (m)	Distance from shore (km)	Water depth (m)
Floating Hywind, NO	2009	1*2.3	65	12	220
Robin Rigg (Solway Firth), UK	2010	60*3	80	9.5	>5
Gunfleet Sands, UK	2010	48*3.6	75	7	2–15
Vänern (Gässlingegrund), SE	2010	10*3	90	4	4–10
Rødsand II, DK	2010	90*2.3	68.5	8.8	6–12
Thanet, UK	2010	100*3	70	12–17.7	14–23
Poseidon, DK	2010	0.011*3			

Source: [16, 17, 18, 19]

SE Sweden, DK Denmark, NL The Netherlands, F Finland, IR Ireland, BE Belgium, DE Germany, NO Norway

added \$0.016–0.017/kWh to wind power projects for the first 10 years of a wind plant's life. Between the middle of 1998 and June 30, 1999, the final day of PTC, more than 800 MW of new wind power generation was installed in the USA, which included between 120 and 250 MW of “repowering” development at several Californian wind power plants. A similar development took place before the end of 2001, which added 1,600 MW between the middle of 2001 and the end of December 2001. Currently, the USA has over 36 GW of installed wind capacity and is leading as world number one in cumulative installed capacity (Table 4 and Table 9). In addition to Texas, major projects have been carried out in the states of Minnesota, Oregon, Wyoming, and Iowa (Table 10).

The typical wind turbine size installed in North America at the end of the 1990s was between 500 and 1,000 kW. The first megawatt turbines were installed in 1999 and since 2001; many projects have used megawatt turbines. At the end of 2009, the typical size was 1,750 kW (2009 editions of [12] and [21]). In comparison to Europe, however, the overall size of wind power plant projects is usually larger. Typical projects in North America are larger than 50 MW, with some projects of up to 200 MW, whereas, in Europe, projects are usually between 20 and 50 MW. The reason for this is the limited space in Europe due to the high population density, especially in Central Europe. These limitations have led to offshore developments in Europe; however, in North America, offshore projects are not a major topic.

Global Wind Power Installations. Table 9 Operational wind power capacity in North America

Country	Installed capacity (MW)			
	End 1995	End 2001	End 2005	End 2010
USA	1,655	4,275	9,149	40,180
Canada	21	198	683	4,009
Total	1,676	4,473	9,832	44,189

Source: 2005 edition of [9] and [13]

The major driver for further wind energy development in several states in the USA are an extension of the PTC as well as fixed quotas combined with green certificate trading, known as the Renewable Portfolio Standard (RPS) in the USA or Renewable Energy Credits (RECs) in the UK. Other drivers are financial incentives, e.g., offered by the California Energy Commission (CEC), as well as green pricing programs. Green Pricing is a marketing program offered by utilities to provide choices for electricity customers to purchase power from environmentally preferred sources. Customers thereby agree to pay higher tariffs for “Green Electricity” and the utilities guarantee to produce the corresponding amount of electricity by using “Green Energy Sources” such as wind energy.

Considering the immense possibilities its wind resources present, Canada is falling somewhat behind the global development of wind energy. Despite this, 2009 marked the best year for Canada's national wind

Global Wind Power Installations. Table 10 Operational wind power capacity in the USA

State	Installed capacity [MW]	
	End 2001	October 2010
Texas	1,100	9,712
Iowa	332	3,669
California	1,688	2,814
Oregon	199	2,095
Washington	161	2,036
Illinois	0	1,847
Minnesota	311	1,813
New York	19	1,274
Colorado	58	1,248
North Dakota	2	1,222
Indiana	0	1,130
Oklahoma	0	1,130
Wyoming	140	1,101
Kansas	114	1,026
Pennsylvania	34	748
New Mexico	1	597
Missouri	0	457
Wisconsin	53	449
West Virginia	414	414
South Dakota	3	412
Montana	0	386
Utah	0	223
Maine	0	200
Idaho	0	163
Nebraska	3	153
Michigan	1	143
Arizona	0	63
Hawaii	11	63
Tennessee	0	29
New Hampshire	25	25
Massachusetts	1	17
Ohio	0	9
Alaska	1	8
New Jersey	8	8

Global Wind Power Installations. Table 10 (Continued)

State	Installed capacity [MW]	
	End 2001	October 2010
Vermont	6	6
Delaware	0	2
Rhode Island	1	1
Total	4,688	36,693

Source: 2002 January edition and 2010 edition of [5]

energy market with 950 MW of new capacity being installed. Since 2007, when the ecoENERGY for Renewable Power Program was established, a payment of C\$0.01/kWh (before tax) is offered for the first 10 years of a project's life. This program was immensely successful, so much so that the target of reaching 4,000 MW of renewable energy projects by March 31, 2011, was achieved well ahead of schedule. As a consequence, the federal government decided to terminate the program in March 2010, and the future remains uncertain. In addition to the federal scheme, the state of Ontario implemented a feed-in tariff, offering C\$0.13/kWh for onshore wind power plants in 2009. With various state governments taking initiative to improve their policy frameworks, substantial growth is expected in the state of Ontario, as well as Quebec, British Colombia, New Brunswick, and Nova Scotia.

Latin America

Despite large wind energy resources in many regions of Latin America, the development of wind-powered electricity is very slow as it can be seen from the figures in Table 11. This is due to the existence of low electricity prices as well as the lack of sufficient wind energy policies. For these reasons, until now, many wind projects in South America have been financially supported by international aid programs.

Argentina, however, introduced a new policy at the end of 1998, which offered financial support to wind energy generation in the form of feed-in tariffs. This initiative triggered a spate of developments until the year 2002, when the economic crisis occurred. With the crisis, most developments stopped, and it was not until 2007 when a new law was established that further development ensued. The federal government in May

Global Wind Power Installations. Table 11 Operational wind power capacity in South and Central America

Country	Installed capacity (MW)		
	End-2001	End-2005	End-2010
Argentina	25.7	27	60
Brazil	20	29	931
Mexico	2.2	3	517
Chile	1.3	2	172
Costa Rica	51	71	123
Other ^a	2.8	80	205
Total	103	212	2,008

Source: 2002 of [5], 2002 edition of [12], 2006 edition of [9], and [13]

^aColombia, Chile, and Cuba

2009 announced the intention to revive the former plan for renewable energy development called GENREN, which will serve as a framework to set up a regulation aiming at the promotion of renewable energies. The finalization and efficient application of this plan in conjunction with the National Wind Energy Strategic Plan (PENEE) is expected to reactivate development in the wind industry, with several projects in the range of 20–90 MW already in the pipeline [22].

In Brazil, the government established the “Programa de Incentivo às Fontes Alternativas de Energia Elétrica” (PROINFA) with the objective to increase the share of renewable energy through government incentives, providing feed-in tariffs and 20 year contracts with guaranteed demand. In addition to this, a number of government tax incentives and discounts on transmission tariffs have also been implemented. The PROINFA program has been an important step for establishing a stable platform for growth in the wind industry, accounting for up to 95% of the current wind power installations in Brazil. Since December 2009, a new approach has been adopted, where bids are now based on production, where the quantity of energy delivered annually is divided by 12 months to smooth the monthly income, rather than on a feed-in tariff system [23]. The policies in Brazil has brought about rapid growth in wind-capacity installations from 341 MW to

606 MW in 2009, a 78.5% increase, and manufacturing bases opening on its shores (2009 edition of [9]).

Mexico has also seen a rapid expansion in wind energy development in recent years, particularly in the Isthmus of Tehuantepec. The average turbine size in this region is 1,136 kW (2009 edition of [12]). However, until now, the only incentive to build new projects has been income tax credits which allow some expenses associated with wind installation to be deducted from taxable income streams.

Chile is another country with favorable conditions for wind energy development. Wind energy is considered by the law as a nonconventional energy resource, which has been promoted by the Development Agency of Chile (CORFO) and the National Commission of Energy (CNE) since 2005. The development of wind energy projects, especially grid-connected electricity-generation projects, is supported through financial incentives. The development of wind parks in the country is still at an early stage however, and there are no specific programs or policies for wind energy alone [22].

Until now, the typical size of wind turbines was around 300 kW in Latin America. However, with new manufacturing capacity for larger wind turbines being established in Brazil, further development is envisioned for the future. In other Latin American countries, however, it may still be difficult to install large turbines due to limitations in infrastructure, particularly for larger equipment such as cranes. Offshore wind projects are not planned, but further small to medium-size (≤ 30 MW) projects are under development onshore.

One of the major incentives for developing wind projects in Latin America is the emission reduction certificate crediting scheme through the Clean Development Mechanism. La Ventosa, for example, the largest wind power plant in Mexico, as well as numerous projects in Chile and Brazil have gained additional income benefits through this scheme.

Asia and Pacific

The Asia-Pacific region contributes to 31% of global installed capacity of wind power (see Table 2), mainly due to the strong growth experienced in India and China. This must be attributed to the rapid growth in demand. Refer to Table 12 for a detailed breakdown of installed capacities in each country.

Global Wind Power Installations. Table 12 Operational wind power capacity in Asia and Pacific

Country	Installed capacity (MW)			
	End-1995	End-2001	End-2005	End-2010
China	44	361	1,266	42,287
India	565	1,426	4,430	13,065
Japan	5	250	1,061	2,304
Australia	10	74	708	1,880
New Zealand	2	37	169	506
Taiwan	0	3	104	519
South Korea	0	8	98	379
Other ^a				98
Total	625	2,162	7,879	61,038

Source: 1997 and 2002 editions of [5], 2006 edition of [9], 2008 edition of [11], and [13]

^aPhilippines, Thailand, Bangladesh, Indonesia, Sri Lanka, Vietnam, and Pacific Islands

India achieved an impressive growth in wind turbine installation in the middle of the 1990s, in what was called the “Indian Boom.” In 1992/1993, the Indian government started to offer special incentives for renewable energy investments by guaranteeing a minimum purchase rate as well as a 100% tax depreciation in the first year of the project. Furthermore, a “power banking” system was introduced, which allowed electricity producers to “bank” their power with the utility and avoid being cut off during times of load shedding. In this scheme, power could be banked for up to 1 year. In addition, some Indian States introduced additional incentives, such as investment subsidies. Development of new installations between 1993 and 1997 were rampant, but it slowed down after 1997 due to uncertainties regarding the future of the incentives (Reference: various editions of [5]). In 2008, the Ministry for New and Renewable Energy (MNRE) announced a national generation-based incentive scheme for grid-connected projects under 49 MW, providing an incentive of 0.5 rupees per kWh (0.7 Euro cents/kWh) in addition to the existing state incentives [24]. This has helped with the development of small- to medium-sized wind power plants, but the tariff is

deemed to be too low to have a significant impact on a project’s viability. In the absence of a better national framework, some states with Renewable Portfolio Standards or other policies to promote wind generation have introduced feed-in-tariffs for wind generation which are higher than that for conventional electricity.

In China, the wind power industry increased capacity by over 100% in 2009 [25]. Its cumulative installed capacity now ranks second in the world. The main reason for the dramatic increase was due to the new focus of the Chinese government to prioritize wind energy as a measure to diversify its energy mix. A recent study by the China Meteorological Administration showed that the potential for wind power in on- and offshore schemes in China are huge, reaching over 3.5 TW in onshore and 200 GW in offshore [25]. The National Energy Administration of China is currently progressing its plans to capitalize on this resource with targets in six provinces to reach a total of 127.5 GW by 2020, dubbed the “Wind Base” program.

There is also the only offshore wind project outside of Europe in China. The Donghai Bridge Wind Power plant is a 102 MW wind power plant close to the Donghai Bridge in Shanghai constituting 34 Sinovel 3-MW turbines. It started producing and transmitting power to the mainland grid on July 6, 2010, and is the first commercial offshore wind power plant in China [26]. As part of the Wind Base program, the Chinese government is planning the construction of further 7 GW of offshore wind power plants in the province of Jiangsu by the year 2020.

In terms of policies, wind power plants have been eligible for a renewable energy premium which is added to the cost of each kWh of electricity sold since 2006. Furthermore, since 2009 a feed-in tariff has been introduced for wind power, which applies to the entire operation period of a wind power plant. The tariff differs depending on the region’s wind resource, and the government is underway to formulate the tariff rate for offshore wind power to be applied to the anticipated growth in the future. Source: 2009 edition of [9].

In Japan, demonstration projects testing different wind turbine technologies dominated the development. At the end of the 1990s, the first commercial wind energy projects started operation on the islands of Hokkaido as well as Okinawa. The interest in wind power is constantly growing in Japan. There also exists

direct financial subsidies aimed at tackling the up-front cost barrier, either for specific equipment or total installed wind system cost.

In Japan and Korea, investment in wind projects are driven by an enhanced feed-in tariff in the form of an explicit monetary reward provided for wind-generated electricity, usually paid by the electricity utility, at a rate per kilowatt-hour somewhat higher than the retail electricity rates being paid by the customer. In addition, these countries employ a form of renewable portfolio standard, which mandates that the electricity utility (often the electricity retailer) source a portion of its electricity supplies from renewable energies.

At the end of the 1990s, the first wind energy projects also materialized in New Zealand and Australia. The main driver for wind energy development in Australia was the Mandatory Renewable Energy Target (MRET) which was introduced in 2001. This scheme requires electricity retailers to source specific proportions of total electricity sales from renewable energy sources according to a fixed timeframe. The national MRET scheme is supported by State-run feed-in-tariff schemes. In addition to the MRET scheme in Australia, there is also a green electricity scheme where customers are given the option to purchase green electricity based on renewable energy from the electric utility at a premium price.

Typical wind turbine sizes installed in the Asia-Pacific region is 1.5–2 MW (Reference: 2009 edition of [12]), and with manufacturing plants based in China and India, further development is anticipated, especially in China.

Middle East and Africa

Wind energy development in Africa is very slow as evident from the figures in Table 13. Most projects require financial support by international aid organizations, as only limited regional support exists. Despite this, an increasing number of African governments are becoming aware of the potential of wind energy in their countries and are beginning to set up the necessary frameworks. An example of this is the setup of the first feed-in tariff in South Africa, which is designed to produce 10 TWh of electricity per year by 2013 from renewable resources [27].

Several projects have been developed in Egypt with the support of the government agency for New and

Global Wind Power Installations. Table 13 Operational wind power capacity in Middle East and Africa

Country/ region	Installed capacity (MW)			
	End- 1995	End- 2001	End- 2005	End- 2010
Egypt	5	125	145	550
Morocco	0	54	64	286
Tunisia		10	20	114
Other	7	14	42	129
Total	12	203	271	1,079

Source: 1997 and 2002 editions of [5], 2006 edition of [9], 2008 edition of [11], and [13]

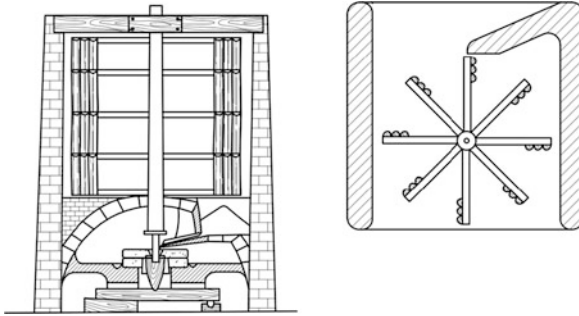
^aIran, South Africa, Cape Verde, Israel, Lebanon, Nigeria, Jordan, Kenya

Renewable Energy Authority (NREA). Five hundred and forty-five megawatt of installed capacity had been built near the city of Zafarana by the end of 2010, in addition to 5 MW in the Gulf of Zayt. Significant developments are expected in the order of 3,000 MW in the next years. Furthermore, there is also a buildup of industrial activities in manufacturing of wind turbines. Further projects are planned in Morocco as well as in Kenya, Ethiopia, Namibia, Tunisia, and Cape Verde. Source: 2009 edition of [11], [22] and various editions of [5].

The typical wind turbine size used in this region is between 600 and 800 kW. However, in light of the fact that the majority of the African population still has no access to electricity grids, small, decentralized, and stand-alone wind energy systems, in combination with other renewable energies, is expected to play a key role. This process is still in its early stage, however, and the limiting factors remain the lack of access to know-how and to adequate financial resources (2009 edition of [11]).

Types of Wind Turbines

Wind energy conversion systems can be divided into those which depend on aerodynamic drag and those which depend on aerodynamic lift. The early Persian (or Chinese) vertical-axis wind wheels (Fig. 1) utilized the drag principle. Drag devices, however, have a very low power coefficient, with a CP_{max} of around ≈ 0.16 [28, 29].



Global Wind Power Installations. Figure 1
Persian-type windmill. (Kaboldy, Wikimedia Commons)



Global Wind Power Installations. Figure 2
Darrieus wind power generator near Heroldstatt, Germany.
(W.Wacker, Wikimedia Commons)

Modern wind turbines are predominately based on aerodynamic lift. Lift devices use aerofoils (blades) that interact with the incoming wind. The force resulting from the aerofoil body intercepting the air flow does not consist only of a drag force component in direction of the flow, but also of a force component that is perpendicular to the drag: the lift force. The lift force is a multiple of the drag force and therefore the relevant driving power of the rotor. By definition, it is perpendicular to the direction of the air flow that is

intercepted by the rotor blade, and via the leverage of the rotor, it causes the necessary driving torque [28–31].

Wind turbines using the aerodynamic lift can be further divided according to the orientation of the spin axis into horizontal axis- and vertical axis-type turbines. Vertical-axis turbines, also known as Darrieus (Fig. 2) after the French engineer who invented it in the 1920s, use vertical, often slightly curved symmetrical aerofoils. Darrieus turbines have the advantage that they operate independently of the wind direction and that the gearbox and generating machinery can be placed at ground level. High torque fluctuations with each revolution, no self-starting capability, as well as limited options for speed regulations in high winds are, however, major disadvantages. Vertical-axis turbines were developed and commercially produced in the 1970s until the end of the 1980s. The largest vertical-axis wind turbine was installed in Canada, the ECOLE C with 4,200 kW. Since the end of the 1980s, however, the research and development of vertical-axis wind turbines has almost stopped worldwide [28, 29, 31, 32].

The horizontal-axis or propeller-type approach dominates the current wind turbine applications. A horizontal-axis wind turbine consists of a tower and a nacelle that is mounted on the top of a tower. The nacelle contains the generator, gearbox, and the rotor. Different mechanisms exist to point the nacelle toward the wind direction or to move the nacelle out of the wind in case of high wind speeds. On small turbines, the rotor and the nacelle are oriented into the wind with a tail vane. On large turbines, the nacelle with rotor is electrically yawed into or out of the wind, in response to a signal from a wind vane.

Horizontal-axis wind turbines typically use a different number of blades, depending on the purpose of the wind turbine. Two or three bladed turbines are usually used for electricity power generation. Turbines with 20 or more blades are used for mechanical water pumping.

The number of rotor blades is indirectly linked to the tip speed ratio, which is the ratio of the blade tip speed and the wind speed. Wind turbines with a high number of blades have a low tip speed ratio but a high starting torque. This high starting torque can be utilized for fully automatically starting water pumping when the wind speed increases. A typical

example for such an application is the water-pumping windmill often seen in the midwest USA. Wind turbines with only two or three blades have a high tip speed ratio, but only a low starting torque. These turbines might need to be started if the wind speed reaches the operation range. But a high tip speed ratio allows the use of a smaller and therefore lighter gearbox to achieve the required high speed at the driving shaft of the power generator [28, 29, 31–33].

Apart from the above discussed wind turbine design philosophies, inventors frequently come up with new designs, using some kind of power augmentation, for instance. However, none of these inventions have given sufficient large-scale performances yet. For the current status of power augmentation wind turbines, see [34] and [35].

The Four Wind Turbine Types

Type-I Type-I wind turbines are fixed-speed wind turbines that were introduced and widely used in the 1980s. It uses a squirrel cage asynchronous generator (SCIG), where the rotor is driven by the turbine and the stator is directly coupled to the grid. The rotation speed of this kind of turbine can vary only slightly, between 1% and 2%; therefore, it is effectively at a “fixed speed.” There are single-speed and double-speed versions, but the double-speed version is better at adapting to different wind speeds, producing less noise at low wind speeds. Type-I wind turbines typically have limited voltage control and reactive power control capabilities to support the grid. The stall system is mostly passive stall, inherit to the aerodynamic design of the blades. There are few options for active control besides connecting and disconnecting, especially if there is no blade pitch change mechanism. However, the concept has been continuously improved, for example, in the active stall designs, where the blade pitch angle can be changed toward stall by the control system. Type-I wind turbines make up 15% of the total cumulative European market share and are manufactured by Suzlon, Nordex, Siemens (used to be Bonus), and Ecotecnia [36].

Type-II Type-II wind turbines were commonly used by Vestas in the 1980s and 1990s in Europe and still offered nowadays by Vestas in selected

markets as well as by Suzlon. They are equipped with a wound rotor induction generator (WRIG), which has limited variable speed capabilities. Power electronics is used to control the rotor’s electrical resistance, which allows both the rotor and the generator to vary their speeds up and down by 10% to cope with wind gusts, to maximize the power quality, and to reduce the mechanical loading on the turbine components, power electronics is also used. Type-II wind turbines also have limited voltage control capabilities to support the grid and are equipped with an active blade pitch control system. Typical Type-II wind turbines are the Vestas models V27, V34, and V47, which make up 5% of the total cumulative European market [36].

Type-III Type-III wind turbines combine the advantages of previous systems with advances in power electronics, producing an improved variable speed capability. It uses a doubly-fed induction generator (DFIG) which has a wound rotor coupled to the grid through a back-to-back voltage source converter that controls the excitation system in order to decouple the mechanical and electrical rotor frequency and to match the grid and rotor frequency. The active and reactive power can be controlled through the use of power electronics, enabling active voltage control. In this type of system, up to approximately 40% of the power output goes directly to the grid, and the window of speed variations is approximately 40% up and down from synchronous speed. Many manufacturers produce Type-III wind generators, including GE, Repower, Vestas, Nordex, Gemasa, Alstom, Acciona Windpower, Suzlon, Bard, and Kenersys. It is the main type of wind turbine in the European market, making up 55% of all installations [36].

Type-IV Type-IV wind turbines are variable speed turbines with full-scale frequency converters that come with the classical drivetrain (geared), in the direct-drive concept (with slow running generator and no gearbox), or in a hybrid version (low step-up gearbox and medium-speed generator). Various types of generators are used for this type of turbine: synchronous generators with wound rotors, permanent magnet generators and squirrel cage induction generators, etc. The stator is connected to the grid via a full-power

electronic converter, and the rotor has excitation windings or permanent magnets. Being completely decoupled from the grid, it can provide an even wider range of operating speeds than the Type-II wind turbine and has a broader range of reactive power and voltage control capacities. Type-IV wind turbines are manufactured by Enercon, MEG (Multibrid), GE, Winwind, Siemens, Leitner, Mtorres and Lagerwey, making up 25% of the total cumulative European market [36].

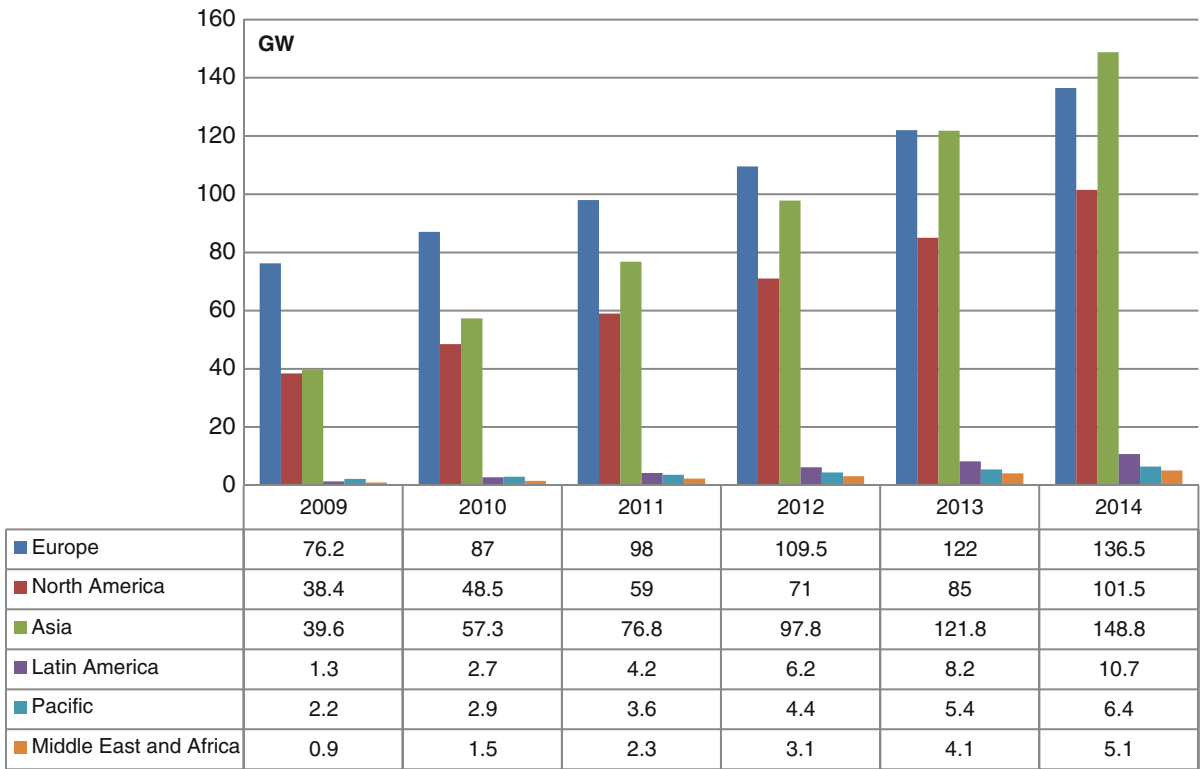
Penetration

At the end of 2009, electricity produced by wind turbines contributed to only 2% of the global electric energy. However, in some countries and regions of the world, wind has become one of the largest electricity sources. The highest shares are seen in countries such as Denmark (20%), Portugal (15%), Spain (14%), and Germany (9%) [2009 edition of [11].

Costs

Over the past 20 years between 1990 and 2010, the capital cost of producing wind turbines has fallen steadily due to economies of scale created by optimization of manufacturing technologies, mass production, and automation [3]. The cost has fallen by a factor of about 4 during the last 25 years between 1985 and 2010, with the general conclusion that costs decrease by some 20% each time the number of units produced doubles.

However, in the years leading up to 2010, particularly since 2006, increasing commodity prices for raw materials such as copper (used in generators), steel (used for towers, gearboxes and rotors), and concrete (used in foundations), rising energy prices, as well as bottlenecks in certain sub-productions have caused the investment cost per MW to increase for new wind power projects (2009 edition of [21]). Although supply chain pressures can be addressed by building new



Global Wind Power Installations. Figure 3
Cumulative market forecast by region 2009–2014 (2009 edition of [9])

production capacity and establishing new manufacturing bases in countries like China, the industry will have to continue battling with the pressure of rising commodity and energy prices.

Despite these challenges, the cost of wind turbine generators has fallen significantly overall, and the industry is recognized as being in the “commercialization phase,” as understood in learning curve theory [37].

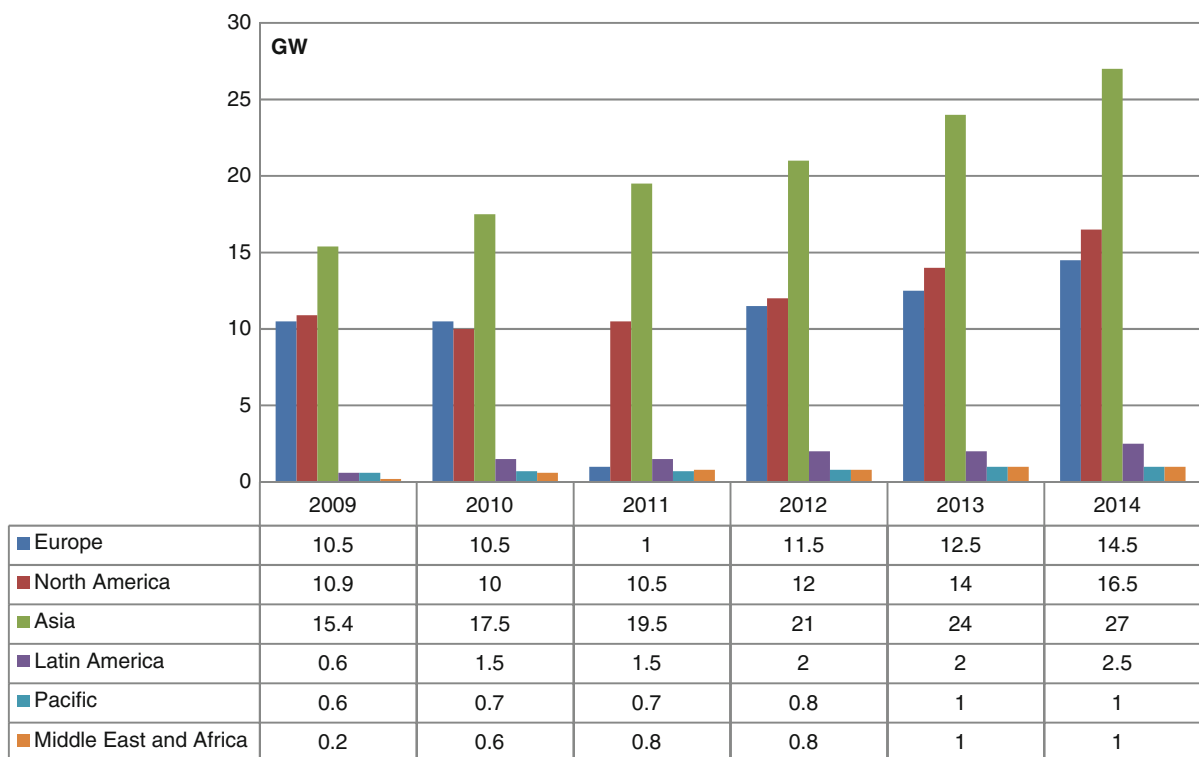
For example, the average capital cost per kilowatt of installed capacity rose from €1,300 in 2007 to €1,350 in 2009. However, it is assumed that it will then fall steadily from 2010 onward to between €1,240 and €1,172 by 2020 and to between €1,216 and €1,093 by 2030 [21].

Given the high up-front costs of wind power projects, large investments of predominantly private but also public funds are expected to flow into the growing wind power markets. This investment will directly benefit regional development by creating jobs in manufacturing, transportation, construction, project development and operation, and maintenance;

providing new revenue sources to local landowners such as a farmers or communities; and increasing the local tax base. The value of investment in the future wind energy markets is thus larger than what is visibly obvious at. Besides these economic benefits, reflected in the economy, wind energy technology also adds value by reducing carbon emission, diversifying fuel supply and providing stable energy production prices. Source: 2009 edition of [12] and [37].

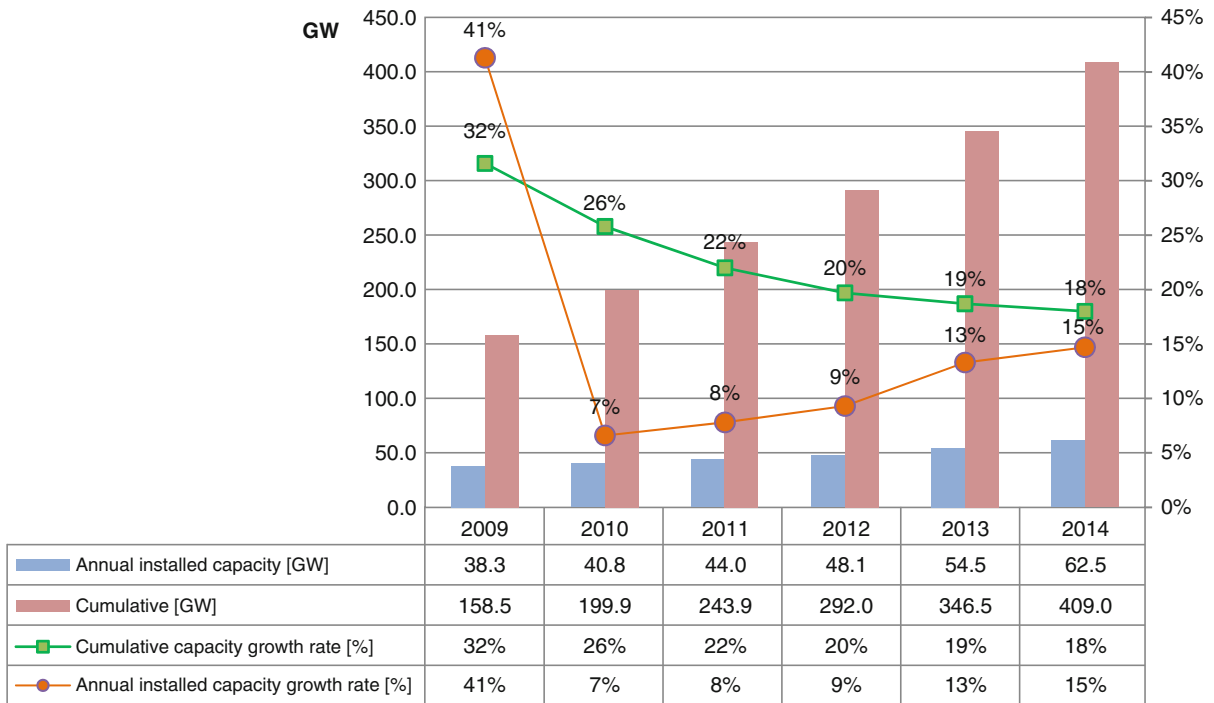
Market Forecast (Future Directions)

2009 saw a 41% market growth in the wind power industry. This was a surprise to many in the field, who expected growth to be staggered by the financial crisis. In particular, growths in the USA and China have been impressive in the past few years, with China single-handedly accounting for one third of the total annual wind-capacity addition in 2009. Although the growth in the USA is expected to stay somewhat flat in the coming years due to the lack of financing and



Global Wind Power Installations. Figure 4

Annual market forecast by region 2009–2014 (2009 edition of [9])



Global Wind Power Installations. Figure 5
Market forecast 2010–2014 (2009 edition of [9])

overall economic downturn, the development in China is set to continue at a breath-taking pace, with well over 20 GW of annual additions. Sustained growth is also expected in India, which will increase its capacity by 2 GW each year. Complimented by growth in other Asian markets, including Japan, Taiwan, South Korea, and the Philippines, the market in Asian is expected to nearly double in the next 5 years, reaching 109 GW of new wind capacity. In Europe, continued growth is also expected, however, not at such a dramatic pace. Regardless, it is expected that a total of 60 GW will be installed by 2014, with large developments in the off-shore market. While Germany and Spain are expected to remain the leading European markets, growth in Italy, France, the UK, and Portugal is expected to continue, as well as opening for some new markets in Poland and Turkey (Fig. 3).

The market in Latin America has also been growing stronger than previously thought. Encouraging developments are seen in Brazil, Mexico, and Chile; however, due to the lack of favorable policies for wind power development in this region, market development will

not reach the heights that would be possible to imagine from the amount of excellent wind resources there is. At the end of 2014, a total of 10.7 GW of installation is expected, increasing from the 1.3 GW in 2009.

In the Pacific region, both Australia and New Zealand seem to be firmly on the track for continued growth at a steady pace, with annual additions of 1 GW by 2014, reaching a total installed capacity of up to 6.4 GW by the end of 2014. Both countries have very good wind resources and a great untapped potential, which is slowly being developed.

In Africa and the Middle East, predictions are less certain, and it is expected that the regions will remain small players in the world's wind market. Substantial plans exist for South Africa however, as well as Kenya, Tanzania, Ethiopia, and Egypt are expected to bring the total capacity up to 5.1 GW by 2014.

As can be seen in Fig. 4 below, in total, it is predicted that in 2014, global wind capacity will stand at 409 GW, with an annual market growth rate of 20.9% being sustained throughout the period in terms of total installed capacity (2009 edition of [9] and [3]) (Fig. 5).

Conclusions

Wind energy has the potential to play an important role in the future energy supply in many areas of the world. Within the last 20 years, wind turbine technology has reached a very reliable and sophisticated level. The growing worldwide market will lead to further improvements, such as large wind turbines or new system applications, including offshore wind power plants. These improvements will lead to further cost reductions and over the medium-term wind energy will not only be able to compete, but be cheaper than conventional fossil fuel power generation technology. Further research, however, will be required in many areas, for example, regarding the network integration of a high penetration of wind energy.

Bibliography

- European Wind Energy Association (EWEA) (2009) Wind Energy the Facts, Brussels, Belgium. <http://www.wind-energy-the-facts.org/>
- Windpower Monthly (2009) 25th Anniversary special – Generating costs – electricity that gets cheaper and cheaper, 01 July 2009. <http://www.windpowermonthly.com/news/972151/25th-Anniversary-Special – Generating-Costs – Electricity-gets-cheaper-cheaper/>
- GWEC (2010) Global wind energy outlook 2010. <http://www.gwec.net/fileadmin/documents/Publications/GWEO%202010%20final.pdf>
- DEWI Magazin (1992) published biannually since 1992 by the German Wind Energy Institute (DEWI). <http://www.dewi.de/dewi/index.php?id=46>
- Windpower Monthly (1985) Industry magazine, monthly, ISSN 0901–7318. <http://www.windpower-monthly.com>
- Gamesa News (2010) Eleven Spanish companies join forces on the Azimut Project to develop a 15-MW offshore wind turbine using 100% Spanish technology. <http://www.gamesacorp.com/en/communication/news/eleven-spanish-companies-join-forces-on-the-azimut-project-to-develop-a-15-mw-offshore-wind-turbine-using-100-spanish-technology.html?idCategoria=0&fechaDesde=&especifica=0&texto=&fechaHasta>
- AEE Wind Power, Wind power observatory, Spanish Wind Energy Association (AEE), Madrid, Spain. http://www.aeeolica.es/en/aee_realizado_informes.php
- IWR International economic platform for renewable energies. <http://www.iwr.de/welcomee.html>
- GWEC Global Wind Report, published by the Global Wind Energy Council (GWEC), Brussels, Belgium. <http://www.gwec.net/>
- BTM Consult Aps. <http://www.btm.dk/>
- WWEA (2010) World Wind Energy Report, World Wind Energy Association (WWEA), Bonn, Germany. <http://www.wwindea.org/home/index.php>
- IEA (2009) Wind Energy Annual Report, International Energy Agency (IEA), Paris, France. <http://www.ieawind.org/>
- GWEC (2010) Global wind statistics 2010. http://www.gwec.net/fileadmin/documents/Publications/GWEC_PRstats_02-02-2011_final.pdf
- Energie Verwertungsagentur (1998) Feed-in tariffs and regulations concerning renewable energy electricity generation in European countries, Vienna, Austria. <http://www.eva.wsr.ac.at/publ/dl.htm>
- IEA (2009) Global renewable energy policies and measures: feed-in tariffs for renewable power (Promotion of Renewable Energy Act). <http://www.iea.org/textbase/pm/?mode=re&action=detail&id=4425>
- EWEA (2009) Operational offshore wind farms in Europe, End 2009. http://www.ewea.org/fileadmin/ewea_documents/documents/statistics/OperationalOffshoreFarms2009.pdf
- 4 C Offshore, Offshore wind farms data base, <http://www.4coffshore.com/>
- Middelgrundens Vindmøllelaug, <http://www.middelgrundens.dk>
- Offshore Wind Energy, The Windenergie-Agentur Bremerhaven/Bremen Magazine, Future Energy Supply from Wind at Sea, WAB. http://www.windenergie-agentur.de/english/PDFs/Offshore_Magazin_Eng2007.pdf
- EWEA (2011) Press release: record 51% growth for EU offshore wind power in 2010. [http://www.ewea.org/index.php?id=60&no_cache=1&tx_ttnews\[tt_news\]=1895&tx_ttnews\[backPid\]=259&cHash=cdd23cd4806718a5dbcf7dbd9f272997](http://www.ewea.org/index.php?id=60&no_cache=1&tx_ttnews[tt_news]=1895&tx_ttnews[backPid]=259&cHash=cdd23cd4806718a5dbcf7dbd9f272997)
- EERE, Wind technologies market report, US Department of Energy, Energy Efficiency and Renewable Energy. <http://eetd.lbl.gov/ea/ems/re-pubs.html>
- TERNA-Studie: Energiewirtschaftliche Rahmenbedingungen und Anreizsysteme für netzgekoppelte Stromproduktion aus erneuerbaren Energien. Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) – German Technical Cooperation, Eschborn, Germany
- McHugh M (2010) Renewable energy investment, acquisitions and business development: Brazil is set to be the new wind energy powerhouse, but which way does the wind really blow? <http://jherrerosdc.typepad.com/jhscd/2010/10/brazil-is-set-to-be-the-new-wind-energy-powerhouse-but-which-way-does-the-wind-really-blow.html>
- GWEC (2009) Indian wind energy outlook 2009. http://www.indianwindpower.com/pdf/GWEO_A4_2008_India_LowRes.pdf
- Junfeng L, Pengfei S, Hu G (2010) China wind power outlook 2010. Chinese Renewable Energy Industries Association, GWEC and Greenpeace. <http://www.greenpeace.org/raw/content/eastasia/press/reports/wind-power-report-english-2010.pdf>
- Feldman S (2009) Solve climate news: China beats US to offshore wind development. <http://solveclimatenews.com/news/20090403/china-beats-us-offshore-wind-development>

27. Gipe P (2009) Renewable energy world, news: South Africa introduces aggressive feed-in tariffs. <http://www.renewable-energyworld.com/rea/news/article/2009/04/south-africa-introduces-aggressive-feed-in-tariffs>
28. Gasch R. Windkraftanlagen. In: Gasch R, Teubner BG (eds) 3 rd edn. Stuttgart, Germany (German version of Wind turbine generators)
29. Gasch R (2002) Wind power plants – fundamentals, design, construction and operation. Solarpraxis, Berlin (Distributed by German Wind Energy Association)
30. Snel H (1998) Review of the present status of rotor aerodynamics. *Wind Energy* 1(S1):46–69
31. Walker JF, Jenkins N (1997) Wind energy technology. Wiley, Chichester
32. Gipe P (1995) Wind energy comes of age. Wiley, New York
33. Thresher RW, Dodge DM (1998) Trends in the evolution of wind turbine generator configurations and systems. *Wind Energy* 1(S1):70–85
34. Van Bussel G (1998) Power augmentation principles for wind turbines. Delft University of Technology, Institute for Wind Energy, the Netherlands. <http://www.ct.tudelft.nl/windenergy/papers/augment/contents.htm>
35. Flay RGJ, Phillips DG, Richards PJ (1999) Development of difuser augmented wind turbine designs in New Zealand. In: Proceedings of the European wind energy conference, Nice, James & James, London, 1–5 Mar 1999, pp 349–352
36. EWEA (2010) Powering Europe: wind energy and the electricity grid. http://www.ewea.org/fileadmin/ewea_documents/documents/publications/reports/Grids_Report_2010.pdf
37. GWEC (2008) Global wind energy outlook 2008. http://www.gwec.net/fileadmin/images/Logos/Corporate/GWEO_A4_2008_lowres.pdf

GM Crop Risk Debate, Science and Socioeconomics

KLAUS AMMANN

Botanical Garden, University of Bern, Bern,
Switzerland

Article Outline

Introduction

Developments in Risk Handling of GM Crops

The Costs and Lost Benefits of Overregulation

The Dispute Between Scientists and Opponents Today

Debate Improvements: What can we do to Enhance the Situation?

Bibliography

Introduction

The General Strategic Situation of the Debate About Green Biotechnology Today

The aim of this text is to set the framework for a better communication about science and regulation, and production of GM crops. GM stands for Genetic Modification, basically an unfortunate denomination, because actually *all* crops are genetically modified, but it is a worldwide accepted term for genetically engineered crops, including transgenes, auto- and allotransgenes, cis- and infra-genes, and synthetic genes, for details see Beardmore [1]. By including gene stacking of various kinds, the situation is getting even more complex [2]. With the introduction of in Vivo Mutation (with Zink-Finger Technology and the latest transformation method transcription activator-like family of type III effectors [TALEs]) the situation will change even more, the age of a high precision and targeted change of genomes has only begun and will develop rapidly, see section [Innovation in Agriculture on All Levels Will Speed Up and Makes it a Necessity to Rethink Regulation Basically and Radically, Most often in the Direction of Lowering the Regulatory Hurdles](#) with details. The term LMOs (Living Modified Organisms), which is generally used in the United Nations Biosafety Protocol (Cartagena Protocol) is nothing but a “Living Proof” that the scientific basis of the Protocol remains questionable, since firstly the term is creating misunderstandings and secondly it is based on an erroneous assumption that GM crops are basically different from conventional crops, as is discussed with detail in the sections [Molecular Processes Similar in Natural Mutation and Transgenesis](#) and [Dissent over Differences Between GM- and Non-GM Crops Causes Transatlantic Regulatory Divide](#). More detailed clarification about the terminology of GMOs is given in a text block of the published Statement of the Pontifical Academy of Sciences: [3].

- There are many different terms used to describe the processes involved in plant breeding. All living organisms are made up of cells in which are contained their genes, which give them their distinctive characteristics. The complete set of genes (the genotype) is encoded in DNA and is referred to as the genome; it is the hereditary information that is passed from parent to

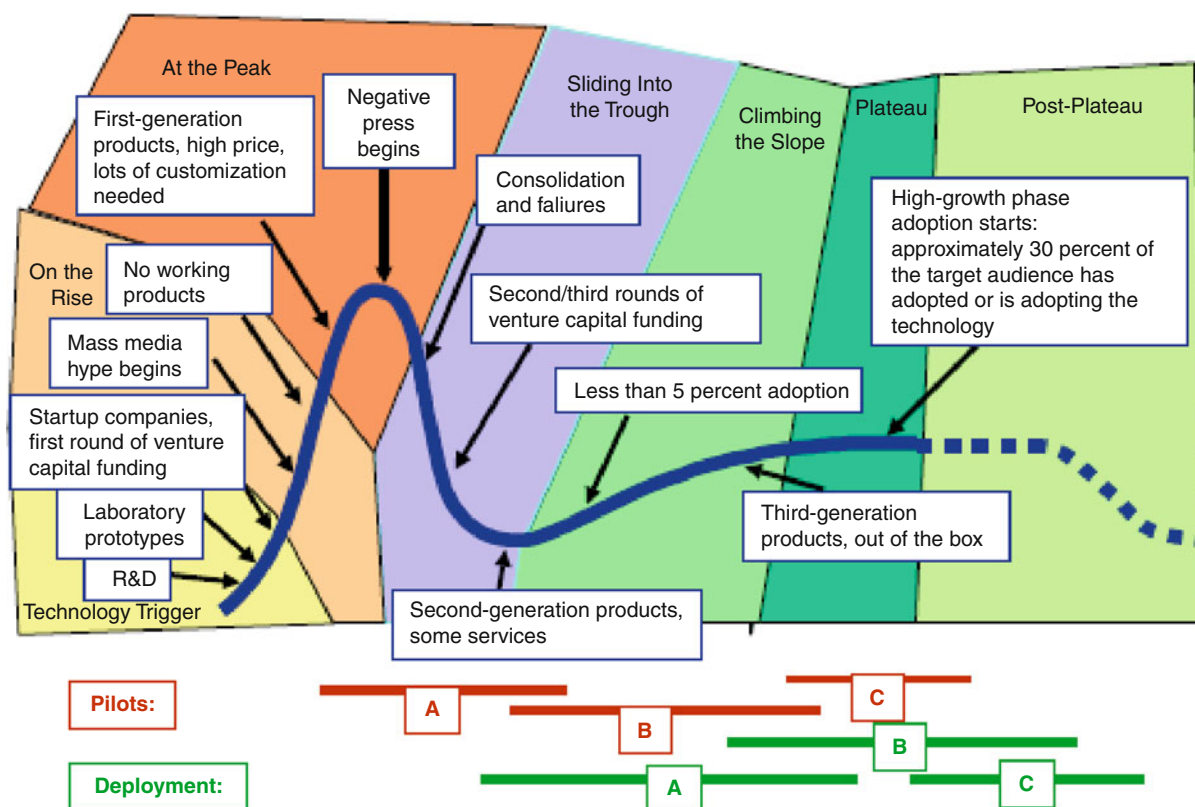
offspring. All plant breeding, and indeed all evolution, involves genetic change or modification followed by selection for beneficial characteristics from among the offspring. Most alterations to a plant's phenotype or observable traits (such as its physical structure, development, biochemical and nutritional properties) result from changes to its genotype. Plant breeding traditionally used the random reshuffling of genes among closely-related and sexually compatible species, often with unpredictable consequences and always with the details of the genetic changes unexplored. In the mid-twentieth century this was supplemented by mutagenesis breeding, the equally random treatment of seeds or whole plants with mutagenic chemicals or high-energy radiation in the hope of generating phenotypic improvements; this, too, gave rise to unpredictable and unexplored genetic consequences from which the plant breeder selected the beneficial traits. Most recently, techniques have been developed allowing the transfer of specific, identified and well characterized genes, or small blocks of genes that confer particular traits, accompanied by a precise analysis of the genetic and phenotypic outcomes: this last category is called 'transgenesis' (because genes are transferred from a donor to a recipient) or 'genetic engineering' (abbreviated to GE in this report) but, in truth, this term applies to all breeding procedures.

The strategic situation in the debate on GM crops is difficult, but not desperate, particularly in Europe – this is an evaluation shared by lots of experts of the debate about agricultural biotechnology; in Europe, it is negatively affecting research and researchers [4]. We have reached in Europe the peak of anxiety related to GM-crops since the introduction of the new technologies, and some opponents to transgenic crops have taken advantage of this situation. They have organized themselves in a veritable protest industry, see section [The Dispute Between Scientists and Opponents Today](#). Nevertheless, the next years should lead to reassurance and scientific consolidation on biotechnology views. We encounter the same repeating dynamics as described for previous technology introductions [5]. The Gartner Hype Cycle [6] adds another dimension to technology life cycle models: it characterizes the typical progression of an emerging technology from user and media overenthusiasm through a period of

disillusionment to an eventual understanding of the technology's relevance and role in a market or domain (Fig. 1).

In the details of the cycle [6], amended by the author – specified for the technology push in transgenic crop development – it should be noted that there are differences between the development of the technologies in the mind of Linden and Fenn and agricultural technologies, where life sciences, combined with regional and cultural diversity, results in a much more diversified picture, often not following the below described phases.

- **2.1 Technology Trigger.** The Technology Trigger is a technological breakthrough, public demonstration, press release or other event that generates significant publicity and industry interest in an emerging technology. Typically no usable products exist, only research and laboratory prototypes (from the first transgenic plants in the 80ties [7]). Venture capitalists may provide some early funding just after the Trigger, if they expect the technology to be a fast runner.
- **2.2 On the Rise.** On the rise to the Peak of Inflated Expectations, media articles explain the technology and discuss its potential impact on business and society. First-generation products emerge like the Flavr-Savr-Tomato [8], but they usually are highly specialized products or extremely difficult to use or with other hitches in the introductory phase. Products are high margin because vendors are still trying to recover R&D costs, and the technology is expensive compared to its cost of production. For example, in 2002, Bluetooth products such as headsets cost \$200, while the final silicon cost of Bluetooth chips likely will be approximately \$5. This is a good stage for venture capitalists to enter the market, before evaluations are at their apex. During this phase, some particularly aggressive enterprises may start to pilot the technology, particularly if it contributes to critical business issues. These enterprises work closely with the vendors to create customized solutions for their requirements
- **2.3 At the Peak of Inflated Expectations.** As the Peak crests, the number of vendors offering the technology increases. These vendors are primarily startup companies and small vendors that try to use the increasing amount of hype for their marketing



Source: Gartner Research (May 2003)

GM Crop Risk Debate, Science and Socioeconomics. Figure 1

Gartner Hype cycle, extended view from [6] after Fig. 3. Technologically aggressive ("Type A") enterprises are relatively comfortable adopting the technology, and moderately aggressive ("Type B") enterprises start to investigate and pilot the technology. Conservative ("Type C") enterprises remain wary (From [6])

benefit. A growing number of enterprises start to examine how the technology may fit within their business strategies, although most do not take action at this stage. Venture capitalists may be interested in selling some of the startups that they equipped with early funding. As problems with first-generation products become visible (e.g., emerging pest resistance in the Bt cotton regions [9, 10] and the latest success message of Huang et al. [11], often because the technology is pushed to its limits, negative publicity starts to push the technology into the Trough of Disillusionment, often the pertinent publications are pushed for negative statements beyond the limit of scientific rules (for example, Web services in 2002 and biometrics in 2003 and two example from the debate on non-target insects related to Bt crops: a) the case of the monarch butterfly [12] and b) Lovei et al. [13] giving false alarm

for ladybirds and its prompt rebuttal by Antony Shelton et al. [14]).

► 2.4 Sliding into the Trough of Disillusionment.

Because the technology does not live up to enterprises' and the media's overinflated expectations, it is rapidly discredited. Some of the early trials end in highly publicized failures. Media interest wanes, except for a few cautionary tales. A significant amount of vendor consolidation and failure occurs. Later-stage investors may be interested in funding vendors during this phase because equity is fairly inexpensive after the "microbubble" at the Peak of Inflated Expectations has burst. However, amid the disillusionment, trials are ongoing and vendors are improving products based on early feedback regarding problems and issues. Some early adopters find some benefit in

adopting the technology. For some slow-moving technologies (for example, biometrics), workable and cost-effective solutions emerge and provide value in niche domains, even while the technology remains in the Trough. The Trough of Disillusionment coincides with the “chasm” in Geoffrey Moore’s classic book, “Crossing the Chasm” [15]. During this stage, vendors need to launch their products from a few early adopters to adoption by a majority of enterprises to begin the climb up the Slope of Enlightenment. There is no real parallel in the GM crop history, except that the differences in GM crop regulation and perception between the Americas and Europe caused a deep transatlantic divide [16].

- ▶ **2.5 Climbing the Slope of Enlightenment.** Focused experimentation and real-world experience by an increasingly diverse range of enterprises lead to a better understanding of the technology’s applicability, risks and benefits. Vendors seek mezzanine or later-round funding for marketing and sales support to pull them-selves up the Slope. Second- and third-generation products are launched by the leading seed companies, and methodologies and tools are added to ease the development process, see the sections under 1.2. The service component declines as a percentage of the sale. Technologically aggressive (“Type A”) enterprises are relatively comfortable adopting the technology, and moderately aggressive (“Type B”) enterprises start to investigate and pilot the technology. Conservative (“Type C”) enterprises remain wary. At the beginning of the slope, the penetration often is significantly less than 5 percent of the potential market segment. This will grow to approximately 30 percent and more as the technology enters the Plateau of Enlightenment. Examples of more or less unexpected enhancements in science and risk assessment of transgenic crops come from a higher precision of gene transfer methods (see sections under 1.2.), also compare to the latest developments in resistance management with a clear success story this year [11].
- ▶ **2.6 Entering the Plateau of Productivity.** The Plateau represents the beginning of mainstream adoption, which began in the Americas much earlier from 2000 onwards, when the real-world benefits of the technology are demonstrated and accepted, see the consecutive reports on the world development of

transgenic crops on www.isaaa.org. Technologies become increasingly embedded into solutions that increasingly are “out of the box,” with decreasing service elements as the technology matures (example conservation tilling). The majority of Type B, then Type C, enterprises adopt the technology. As a high-profile technology matures, an “ecosystem” often evolves around it. The ecosystem supports multiple providers of products and services, and also a market for related products and services that extend or are based on the technology (for example, virtual private networks in 2003 or the growing market for suppliers of molecular laboratories or the growing market for electronic equipment for precision agriculture).

The final height of the Plateau varies according to whether the technology is broadly applicable or benefits only a niche market, depending heavily on crop and region.

- ▶ **2.7 Post-Plateau.** As a technology achieves full maturity and supports thousands of enterprises and millions of users, producers and consumers, its hype typically disappears, as seen in the Americas. Only a few specialist magazines continue coverage of new aspects of implementing and maintaining the technology. Often there may be innovations around this technology that will follow their own Hype Cycles (new crop varieties on stress resistance, on bio-fortification, pharmaceutical crop lines etc.).
- ▶ **3.0 The Time-to-Maturity Assessment.** Technologies do not move at a uniform speed through the Hype Cycle. It often takes years for a technology to traverse the Hype Cycle — some technologies like GM crops may take decades, with considerable regional differences. There are three adoption speeds: “Fast-track” technologies go through the Hype Cycle within two to four years. This occurs when the performance curve inflects early in the life cycle of a technology. These technologies find themselves adopted without much fanfare, bypassing the Peak of Inflated Expectations and Trough of Disillusionment. Many enterprises are unaware of their sudden maturity and applicability, such as what has happened with instant messaging and Short Message Service.

It is interesting to note that the Showalter “hystories” on the introduction of most new technologies [5]

report no real damage in their subsequent introductory phase, or the benefits were so overwhelming that the debate was soon fading away. This alone demonstrates clearly that it is the sociocultural environment strongly influencing the risk debate [17]. The most recent events seem to hint that Europe finally finds to a more decontracted way of looking at GM crops: The new report of the Royal Society [18] tries to unite conventional and biotechnology approaches for the sake of making progress on agricultural management in developing countries:

- Past debates about agricultural technology have tended to involve different parties arguing for either advanced biotechnology including GM, improved conventional agricultural practice or low-input methods. We do not consider that these approaches are mutually exclusive: improvements to all systems require high-quality science. Global food insecurity is the product of a set of interrelated local problems of food production and consumption. The diversity of these problems needs to be reflected in the diversity of scientific approaches used to tackle them. Rather than focusing on particular scientific tools and techniques, the approaches should be evaluated in terms of their outcomes.

It might well be that we arrive sooner than expected from a period of disillusionment to an eventual understanding of the technology's relevance and role in a market or domain.

Innovation in Agriculture on all Levels will Speed up and Makes it a Necessity to Rethink Regulation Basically and Radically, most Often in the Direction of Lowering the Regulatory Hurdles

Unfortunately, regulatory legislation is in its nature static, needs a long time to be settled in international negotiations, and then, finally, settled and approved with an important number of signatory states as the Cartagena Protocol; therefore, it is nearly impossible to make the necessary changes based on good science. At the time of the establishment of the Cartagena Biosafety Protocol, the similarities between nontransgenic and transgenic organisms on the molecular level were not widely known, although properly published (see latest review with early publications [19]), and

a correction about these grave errors (recently called by the author as “Genomic Misconception,” publication in preparation) in concept is now nearly impossible – details in section [GM- and Non-GM-Crop Differences Over-Estimated](#), the “[Genomic Misconception](#)”. But the situation is not getting better: the accelerating speed of scientific progress and discoveries used for new (agricultural) technologies is breathtaking. A short overview is provided in the following sections.

New Biotechnology Approaches in Plant Breeding,

Introduction In an early paper, Britt et al. give an overview on many molecular possibilities which will develop for new breeding successes [20], they address the current status of plant gene targeting and what is known about the associated plant DNA repair mechanisms. One of the greatest hurdle that plant biologists face in assigning gene function and in crop improvement is the lack of efficient and robust technologies to generate gene replacements or targeted gene knock-outs. They also face an old problem in plant breeding summarized under the complex term of epigenetics [21, 22], a problem corrected in conventional plant breeding by careful and often tedious selection processes. Unfortunately, opponents abuse epigenetics as a seemingly new problem for genetic engineering [23], avoiding the mention of modern molecular insight and its ease to correct such problems in a more targeted way. It is clear that transgenesis will remain a solid technology for breeding, but new approaches will appear – as science is always open for progress and new breakthroughs. Here, we only mention shortly progress from another more holistic perspective of systems biology: the dynamics of Metabolomics [24], and also the growing speed of discovery in proteomics [25], techniques which will increasingly augment more common types of experimentation, especially as they provide the capacity of generating data sets that can be compared across studies and laboratories [26], and because quantitative proteomics data are generated with unprecedented sensitivity, accuracy, and reproducibility. There are many new biotechnologies enhancing the speed of achieving targeted breeding successes such as the high throughput marker finding technology [27, 28], only a few can be mentioned here:

Cis- and Intragenic Approaches A new technology has now proven to be a successful strategy: As Romments describe it, cisgenetics is a welcome way of combining the benefits of traditional breeding with modern biotechnology. It is an understandable enthusiasm of the first researchers using this technology to emphasize the positive sides by also comparing to transgenesis as an “old-fashioned” method with its problems. But things are certainly not so easy: In sections [Molecular Processes Similar in Natural Mutation and Transgenesis](#) and [Dissent Over Differences Between GM- and Non-GM Crops Causes Transatlantic Regulatory Divide](#), it is made clear that on the genomic level, particularly on the level of molecular processes, there is no difference between transgenic and nontransgenic crops (supported by an important body of scientific literature), and this is certainly also true to cisgenic and intragenic varieties. This is why it is questionable and based on false grounds to make claims that those new methods in transformation would be safer, as Giddings has made it clear in his letter [29], and his arguments against the views of [30–32] and later publications [33–35] could have been targeted as well: they try to demonstrate that the new cisgenics and intragenics are safer than transgenics, which is not based on any facts, rather it is based on accepting without scientific scrutiny the negative public perception on transgenic crops. It is also wrong to use without clarification the term “alien genes” in view of confirmed and widely accepted universality of DNA and genomic structures.

However, there is nothing to say against the application of such new methods per se, as [33, 34] can demonstrate:

- The classical methods of alien gene transfer by traditional breeding yielded fruitful results. However, modern varieties demand a growing number of combined traits, for which pre-breeding methods with wild species are often needed. Introgression and translocation breeding require time consuming backcrosses and simultaneous selection steps to overcome linkage drag. Breeding of crops using the traditional sources of genetic variation by cisgenetics can speed up the whole process dramatically, along with usage of existing promising varieties. This is specifically the case with complex (allo)polyploids and with

heterozygous, vegetative propagated crops. Therefore, we believe that cisgenetics is the basis of the second/ever green revolution needed in traditional plant breeding. For this goal to be achieved, exemption of the GM-regulation of cisgenes is needed.

Reverse Screening Methods: Tilling and Eco-Tilling

Two rather independent publications [36, 37] with largely incongruent literature lists promote a new technology of finding useful genes within the genome of the crops involved: They both promote powerful reverse genetic strategies that allow the detection of induced point mutations in individuals of the mutagenized populations, can address the major challenge of linking sequence information to the biological function of genes, and can also identify novel variation for plant breeding [37]. Rigola et al. [36] develop reverse genetics approaches which rely on the detection of sequence alterations in target genes to identify allelic variants among mutant or natural populations. Current (pre-) screening methods such as *tilling* and *eco-tilling* are based on the detection of single base mismatches in heteroduplexes using endonucleases such as CEL 1. However, there are drawbacks in the use of endonucleases due to their relatively poor cleavage efficiency and exonuclease activity. Moreover, prescreening methods do not reveal information about the nature of sequence changes and their possible impact on gene function. Rigola et al. [36] present a *KeyPointTM* technology, a high-throughput mutation/polymorphism discovery technique based on massive parallel sequencing of target genes amplified from mutant or natural populations. Thus, *KeyPointTM* combines multi-dimensional pooling of large numbers of individual DNA samples and the use of sample identification tags (“sample barcoding”) with next-generation sequencing technology. Rigola et al. [36] can demonstrate first successes in tomato breeding by identifying two mutants in the tomato *eIF4E* gene based on screening more than 3,000 M2 families in a single GS FLX sequencing run, and discovery of six haplotypes of tomato *eIF4E* gene by re-sequencing three amplicons in a subset of 92 tomato lines from the EU-SOL core collection. This technology will prove to be useful and does not need for its own breakthrough to refer to a scientifically unjustified critique of transgenesis.

Whether the new technology will replace the transgenic “Amflora potato” has still to be proven by further scrutinizing of the results of the equivalent trait [38].

Zinc Finger Targeted Insertion of Transgenes Plant breeding has gone through dynamic developments, from marker-assisted breeding to transgenesis with steadily improved methods to the latest development of the Zinc-finger enzyme-assisted targeted insertion of transgenes in complex organisms [39–42]. Zinc-finger nucleases (ZFNs) allow gene editing in live cells by inducing a targeted DNA double-strand break (DSB) at a specific genomic locus. However, strategies for characterizing the genome-wide specificity of ZFNs remain limited. According to [43], comprehensive mapping of ZFN activity in vivo will facilitate the broad application of these reagents in translational research.

The development toward more insertion precision and less genomic disturbance is so rapid that promoters of organic farming will see dwindling one of their pet arguments even more rapidly: Genomic disturbance of modern breeding is certainly less important and will even be negligible compared to the old breeding methods, still promoted stubbornly by the organic plant breeding community [44]: It is very likely that the transcriptomic disturbances will be even smaller in future – compared to the clumsy and tedious methods of conventional breeding, see also the latest developments in sections **TALEs: Transformation Method Transcription Activator-like Family of Type III Effectors** and **Precision Engineering Through DNE Meganucleases** below.

TALEs: Transformation Method Transcription Activator-like Family of Type III Effectors The generation of double-strand DNA breaks (DSBs) promotes homologous recombination in eukaryotes and can facilitate gene targeting, additions, deletions, and inactivation. Zinc-finger nucleases have been used to generate DSBs and subsequently for genome editing, but with low efficiency and reproducibility. In contrast, the transcription activator-like family of type III effectors (TALEs) contains a central domain of tandem repeats that could be engineered to bind specific DNA targets. The new method is capable of generating site-specific DSBs and has great potential for site-specific

genome modification in plants and eukaryotes in general [45]. See also comments on the newswire CNBS [46] on the discovery:

- Dr. Mahfouz has developed a “repair tool” (molecular scissors) made out of protein that does two things: it finds the exact place on the genome where it is to be cut using a genetic “postcode” and then deletes, adds or edits the gene with great accuracy and precision.

Dr. Mahfouz’s work has the potential for much broader applications including human health. This new technology could enhance the technique that may be used to substitute “good” genes for bad, or to cut out or silence the defective genes that cause disease.

Commenting on the research, KAUST Provost Stefan Catsicas saw the technology as a scientific breakthrough and, if the patent is eventually successful, having potentially promising revenues. Dr. Nina Fedoroff, Professor of the Life Sciences at Penn State University, said the Mahfouz paper “shows the practicability of creating DNA-cutting enzymes tailored to cut a desired target sequence with very high specificity. This is an excellent step forward toward creating very specific genetic improvements in crop plants, while avoiding the potential risks many are concerned about with more conventional genetic modification strategies. Moreover, the paper gives the first evidence that this particular strategy will work in plants.” Professor Fedoroff is “delighted to see such cutting-edge contributions emerging from a university as young as KAUST!”.

Precision Engineering Through DNE Meganucleases

Engineered DNE meganucleases can be used for cloning and molecular analysis purposes in much the same ways as conventional restriction enzymes. The important difference, of course, is that meganucleases recognize much rarer DNA sequences than restriction enzymes. This makes them particularly well suited to the manipulation of extremely large DNA sequences such as intact genomes. Importantly, DNE meganucleases cleave to leave four base pair 3’ overhangs suitable for “sticky-end” cloning. The first application with a new tool called Directed Nuclease Editor™ in plant breeding by Bayer Crop Science <http://www.precisionbiosciences.com/> seems promising: The meganucleases have been first

used to do precision work in human gene therapy, but an outlook into various other applications was announced as early as 2003 [47–49].

Synthetic Biology In some 150 laboratories, synthetic biology is intensively researched, and it seems clear that the future will bring here some unexpected revolutions: A new field, synthetic biology, is emerging on the basis of these experiments [50], where chemistry mimics biological processes as complicated as Darwinian evolution. According to [51], the emerging field of synthetic biology is generating insatiable demands for synthetic genes, which far exceed existing gene synthesis capabilities. Tian et al. claim that technologies and trends potentially will lead to breakthroughs in the development of accurate, low-cost, and high-throughput gene synthesis technology – the capability of generating unlimited supplies of DNA molecules of any sequence or size will transform biomedical and any biotechnology research in the near future. And, according to [52], already in 1998 the redesigning of nucleic acids has been judged in an optimistic way, this was confirmed in an important Nature review in 2005 [53].

The real breakthrough came with the synthesis of an organism including its reproduction, achieved after years of research and a firm belief in success, typical of the senior author of the mega project still continuing, [54–57].

A pragmatic view of a new regulatory scheme answering the new biosafety tasks of synthetic biology is proposed by [58] (Fig. 2):

This kind of new regulatory approach will be necessary in order to avoid unnecessary hindering of research progress in synthetic biology, a demand supported with other innovative suggestions for interactive procedures [59]. Another balanced view [60] demonstrates also the new risks arising from synthetic organisms and the accidental (or purposeful) release in the environment. As always, the ethical awareness and behavior has to be developed further, agreeing with [61] not in a way which gives forfeit power to social sciences. What we really need is a new interfaculty, interdisciplinary or, even better, transdisciplinary discursive scheme as proposed in sections [Long Term Discourse and Decision Making Processes](#) and [The Second Generation Systems Approach as a New Decision Making Process](#).

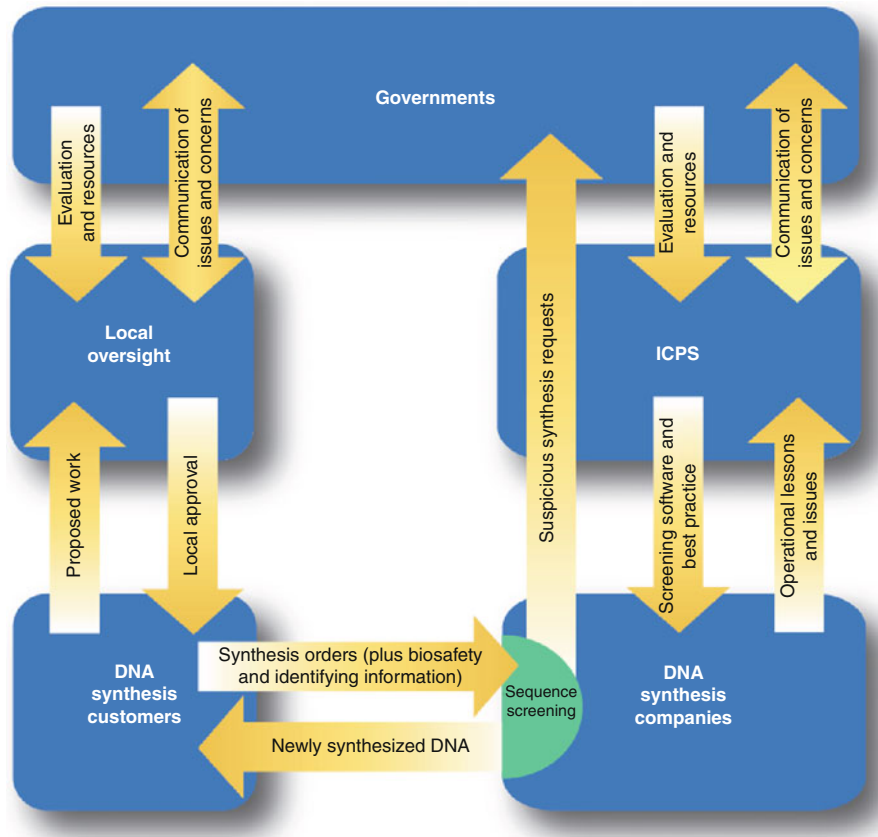
What happened some 35 years ago in the US National Institute of Health in the words of Henry I. Miller [62] should be a warning.

- Thirty-five years ago, the US National Institutes of Health adopted overly risk-averse guidelines for research using recombinant DNA, or “genetic engineering,” techniques. Those guidelines, based on what has proved to be an idiosyncratic and largely invalid set of assumptions, sent a powerful message that scientists and the federal government were taking seriously speculative, exaggerated risk scenarios – a message that has afflicted the technology’s development worldwide ever since.

A final remark: In a way, the artificial altering of genes producing Bt toxins can, strictly spoken, also be summarized under synthetic biology since the specifically altered Bt toxins in order to facilitate resistance management of Bt crops: Bruce Tabashnik, who works on problem solving programs for Bt crops with field research and new concepts of resistance management [63]: Relative to native toxins, the potency of modified toxins was >350-fold higher against resistant strains of *Plutella xylostella* and *Ostrinia nubilalis*. Previous results suggested that the modified toxins would be effective only if resistance was linked with mutations in genes encoding toxin-binding cadherin proteins [64]. Tabashnik et al. report evidence from five major crop pests refuting the Soberon hypothesis.

Illusions and Realities on Educational Effects in the Debate, the Dialogue Between Science and the Public

There is no doubt that there is hope and need to simply start and/or maintain an open dialogue between major stakeholders among young scientists, politicians, industry, and society [65], although there are many obstacles such as asymmetric relationships among the partners, which can render the discourse complex and unpredictable. And it is uncontested here that education on all school levels has its justified place; this has again been shown with empirical results from Spain [66, 67]. Gensuisse should also be mentioned here with educational activities in schools and a popular open day of Genetics in major Swiss cities organized by researchers and institutes every year [68]. And education on biotechnology in the developing world is especially important,



GM Crop Risk Debate, Science and Socioeconomics. Figure 2

Our framework calls for the immediate and systematic implementation of a tiered DNA synthesis order screening process. To promote and establish accountability, individuals who place orders for DNA synthesis would be required to identify themselves, their home organization, and all relevant biosafety information. Next, individual companies would use validated software tools to check synthesis orders against a set of select agents or sequences to help ensure regulatory compliance and flag synthesis orders for further review. Finally, DNA synthesis and synthetic biology companies would work together through the ICPS, and interface with appropriate government agencies (worldwide), to rapidly and continually improve the underlying technologies used to screen orders and identify potentially dangerous sequences, as well as develop a clearly defined process to report behavior that falls outside of the agreed-upon guidelines. ICPS, International Consortium for Polynucleotide Synthesis (From [58])

if done in a participative way, and with proper ramifications in all institutions of communication, science, and regulation: In April 2007, biosafety and biotechnology scientists, regulators, educators, and communicators from Kenya, Tanzania, and Uganda met to examine the status and needs of biosafety training and educational programs in East Africa [69].

Thus, educational efforts on all levels are not in vain, and deplorably there are too few academic institutions active in biotechnology education [70].

The structure of the debate has shifted: Today, the GM crop debate is steered by scientific *and* pseudoscientific arguments. And this also includes an element of hope for the pro-scene: Slowly but surely the pseudoscientific arguments are fading away for the opponents, since there is no serious incident known despite the fact that millions of hectares are grown with GM crops worldwide [71].

There is a widespread mistrust against new technologies where everybody feels it will change their own life,

and this often happens in a phase where the benefits are not yet clearly visible, especially for the consumers/users. But it is not correct to reduce those difficulties to an exclusive criticism of the so-called deficit model [72–74] where the people just have to be educated and then they would refrain from negative emotions. A question mark on the exclusive use of the “deficit model” is justified, but surprising conclusions emerge from the above-mentioned critics themselves: They do not discard altogether the traditional deficit model, rather they propose to combine it with the *contextual approach*, thus emphasizing the complex and interacting nature of the knowledge-attitude interface. This highlights the sophistication and value of lay understandings of science that can exist in the absence of formal scientific knowledge [75, 76]. Surprisingly, positive are results of polls which are conducted by Philip Aerni with more closeness to the real life and careful avoiding of polling mistakes [77], the study concludes:

- The results of our discrete choice analysis show that Swiss consumers treat GM foods just like any other type of novel food. We conclude from our findings that consumers tend to appreciate transparency and freedom of choice even if one of the offered product types is labeled as containing a genetically modified ingredient. Retailers should allow consumers to make their own choice and accept the fact that not all people appear to be afraid of GM food. [77]

There is growing consensus that scientific knowledge extends beyond the simple learning of “facts” that can be straightforwardly defined and measured [78]. From this perspective, privileging formal scientific knowledge as the sole basis of rational preference formation leads us to overlook other knowledge domains that may be equal or even more important determinants of attitudes toward science.

These insights have been condensed into a feasible discursive method of the *Systems Approach* initiated by Churchman [79] and refined by Rittel et al. [80–82]. Details on the methodology are given under sections [Long Term Discourse and Decision Making Processes](#) and [The Second Generation Systems Approach as a New Decision Making Process](#), where the *solutions* are discussed.

It is an illusion to solve ill-fated GM-disputes by just adding social and cultural aspects, or that the dispute should, so to say, start from the other end of the controversy ignoring the biosafety science [83] or even worse to primarily appeal to feelings and emotions of the public and indulge in entertaining but ultimately meaningless discussions in order to catch the interest of the public – we should not mimic the strategy of the protest corporations. That said, this does not mean that sociocultural aspects including emotions should be neglected – even the boulevard press sends out strong signals for learning processes. Vaughan’s [84] plea is that regulatory officials should engage in an interactive process of information and opinion exchange that is reasonable and effective within vastly different socioeconomic and cultural contexts. This is often a challenge to government employees concentrating on office work routine. Patricia Osseweijer [85, 86] offers an interesting compromise: a mix of science, ethics, and emotions with her “Three E-Model” Entertainment (getting attention), Emotion (identification), and Education (information and skills for [future] decision making). It has been developed on the basis of long-term experience and observation of public communication by individuals in the Department of Biotechnology of the Delft University of Technology [87, 88].

Despite all possible refinements and enhancements in the dialogue with the public, we should not underestimate the negative role of the opponents of genetic engineering in plant breeding organized as professional protest corporations, see section [The Costs and Loss Benefits of Overregulation](#).

How the Internet is Influencing the Debate The Internet as a worldwide literacy practice environment is still underestimated, nevertheless it has created a new situation in communication, providing a new dynamic field for research and knowledge accumulation [89]. It has created an Internet-based debate culture with all its ramifications from classic email over blogs and better organized social media to twitter and this not only in nanotechnology [90], but also in other research realms and E-business [91]. The evolution in this kind of debate is still going on with unprecedented dynamics and is not yet fully understood in all its consequences [92], [93], and [94]. The hope is that easier communication through the Internet will invite a *collaborative*

instead of *confronting* modus [95]. Some advice on how to behave in chats and blog debates on the Internet might be useful [96]; compare a list of useful websites and databases on biosafety by DeGrassi et al. [97] and [98]. A list of pertinent websites can be expanded ad libitum, the present state of error of 2011, with all the personal bias in [99].

Informatics and the new ease to access huge amounts of scientific information on the Internet causes a democratization effect on the science debate. But this can only then lead to positive developments if the new flood of information is also well organized and provided people make serious efforts to analyze the available information, so that our understanding of complex scientific knowledge can indeed be improved. As Janetzko (2008) shows, it is not enough to make use of the most common search machines, only professionally organized searches and databases on scientific literature can help and create some limited reliability and sustainability of scientific knowledge. And: clearly, the usual citation clusters among opinion-buddies will not suffice. And it should be emphasized: Electronic ease does not replace the tough job of scholarly reading and understanding. It will be a difficult task for the future to divide up clever knowledge accumulation and genuine thinking work among active scientists. A caveat already signaled by Seneca: Thoughtful Action creates more wisdom than knowledge accumulation, can be interpreted related to social electronic networking in two ways: On one side, the immense intensification of social networking via the Internet creates among other things a new possibility for post-publication reviewing and filtering out the really relevant publications and ideas. On the other hand, it hinders systematically the deepening of your own knowledge in an individual way, and be it only by reading every year a dozen or two really relevant book publications.

This major shift from paper to electronics is also creating new methods of *quantitative* analysis of scientific work: see the Scientometrics Wikipedia: <http://en.wikipedia.org/wiki/Scientometrics>. Actually, this newly emerging science can provide with caveats and insights into changes in research priorities, reveal citation habits, evaluate journals with new scales, etc. [100–103]. A typical example is given in the analysis of the coming and going of the Frankenfood myth [104], with a somewhat surprisingly early and sharp

peak of appearances of the word Frankenfood in websites for 1998, followed by a sharp decline to virtually zero 2 years later (Fig. 3).

This figure is confirmed in [104] with the following statements and figures (Figs. 4 and 5):

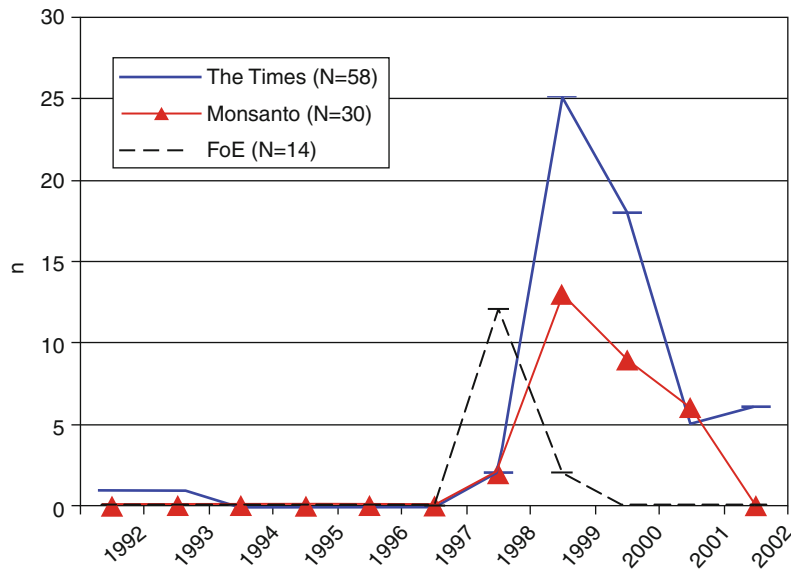
The comments in [104]:

- Our interpretation of these results is as follows: the decline of the organizing power of the metaphor was rapid in 1999 and 2000 when the metaphors of 'Frankenfood' and 'Frankenstein food' began to be outdated. Due to its generalized meaning, the metaphor was used increasingly across domains and therefore lost its domain-specificity and the ability to organize distinctions among domains. This might also explain why the NGOs stopped using the metaphor in 2000 (HELLSTEN, 2003). From [104]

Scientometrics can do much more, [105] have shown the potential of a sophisticated statistical analysis combined with modeling of community interactions in the web: Besides tracking just the description-to-acquisition behavior of users, scientometrics can do much more by longer observation periods which offers the chance to make richer inferences about both group and individual user intentions – trends of intruding into human behavior and making conclusions, which are actually beyond Orwell's imagination. Yet we should have no illusions, since a lot of work and application is already going on in the marketing and advertisement scene, which has also an often manifested interest in knowledge accumulation methods [106, 107]. It is somehow amazing to realize that the academic world in most fields of specialization have not yet reached the realms of professional knowledge accumulation and consolidation – not to speak about an efficient way of reaching out from knowledge accumulation to efficient development of new technology. Scientometrics would have the potential to get instrumentalized in research and development, with some good chance to be used also in new peer review processes.

A *qualitative* evaluation of science should involve additional elements – see below under peer review in the section [Developments in Risk Handling of GM Crops](#) on regulation.

Deplorably, important networks are often only known in specific reader clusters, these awareness gaps



GM Crop Risk Debate, Science and Socioeconomics. Figure 3

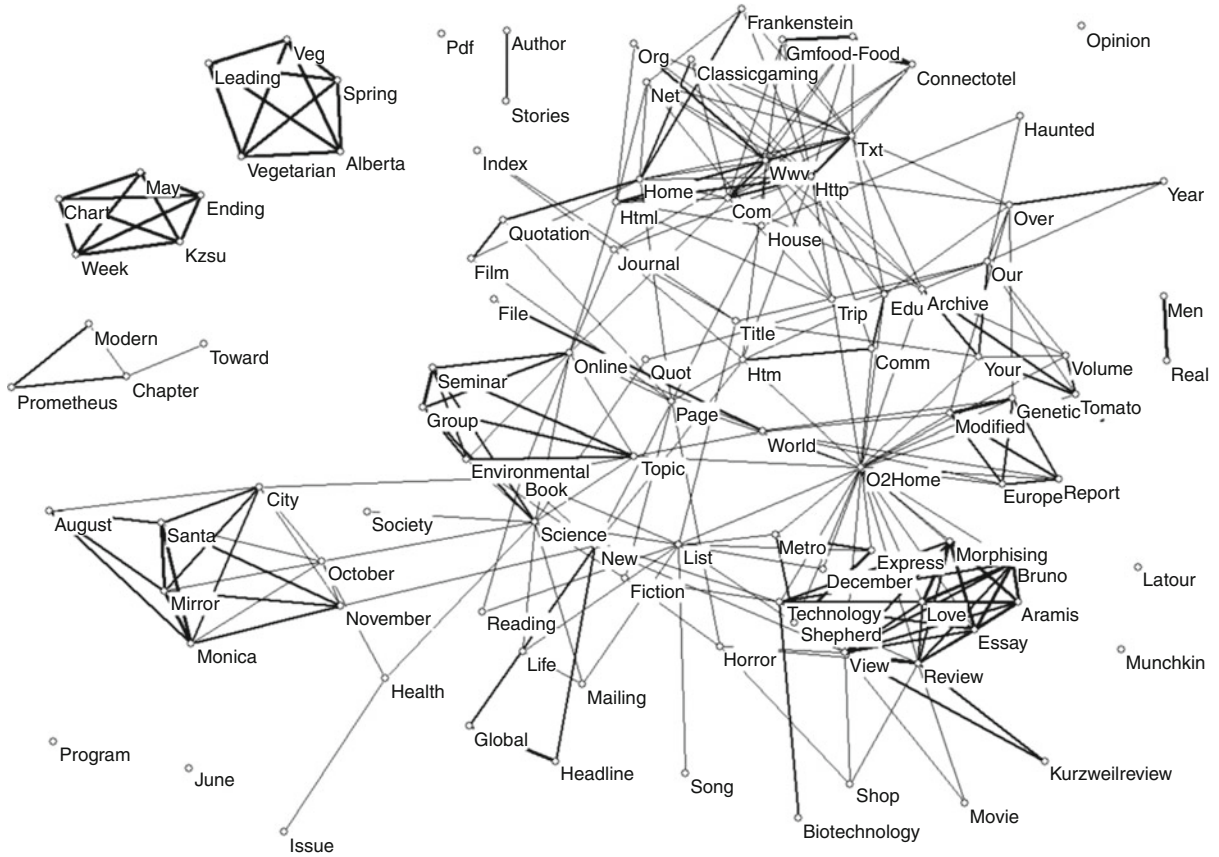
Web site pages addressing the “Frankenfood” and “Frankenstein food” issues at Monsanto, the *Times*, and the Friends of the Earth Web sites. jcmc.indiana.edu/vol8/issue4/hellsten.html

should be minimized. We need knowledge exchange, jumping over geographical and ideology fences.

Science Education and New Developments on the Internet In a successful initiative, Ron LaPorte and his group “Supercourse” started in 2002 [108] a new educational Internet-based system: In his view, Journals do not have an exclusive “right” to science. A publication and a scientific presentation do virtually the same thing – they share scientific knowledge. Publication and presentation have been separate but could “morph” into a single entity. This metamorphosis is taking place and is driven by a juggernaut called PowerPoint, Microsoft’s graphics and slide presentation software, and today enriched with more media from Twitter over YouTube to all the numerous blog systems, networking enhanced with RSS, etc. More on the Supercourse program in [109–113], also in connection with the Bibliotheca Alexandrina in Egypt: [114] Another possibility on a well-organized collection of Powerpoint slides is offered for free by the University of California by Peggy Lemaux and Barbara Alonso, University of California http://ucbiotech.org/resources/slide_archive/index.html. A series of over 100 slideshows is offered by the bibliography of the

author; new slide shows are continuously added, they can be downloaded from [115]. An important new development started 2002 at the Bibliotheca Alexandrina, where a new world center of electronic knowledge is emerging, which is based on thoughtful new structures [116].

On Biosafety Education Biosafety is today a permanent topic on local, national, and international level, and basically, it is good to see educational activity. As demonstrated in this contribution, the topic of biosafety is highly controversial, and so are the views on the various educational activities. The most blatant misunderstandings in biosafety education stem again from the “Genomic Misconception,” which forces authors seemingly to focus on transgenic crops alone, which is scientifically unacceptable as we will see in section **GM- and Non-GM-crop Differences Overestimated, the “Genomic Misconception”**. A symptomatic example on the enumeration of risks related to transgenic crops is given by Craig et al. [117]: All risks duly mentioned can be attributed just as well to conventional crops. The only difference between modern and conventional breeds can be found in risk mitigation, which is much easier in the case of the transgenic



GM Crop Risk Debate, Science and Socioeconomics. Figure 4

The cosine map of 107 words used more than once in the 205 documents on Frankenfoods in 1999 (cosine ≥ 0.1)
(From [104])

crops. Here, just two recent examples related to the successful prevention of upcoming resistant pest insects (a problem arising in all kinds of agricultural management systems): [63] and [11]. It is deplorable, that most biosafety education is still based on the erroneous “Genomic Misconception,” which results automatically into a biosafety risk view focusing on the process of transgenesis instead of working on a product-oriented basis. More about the “Genomic Misconception” is discussed in section [GM- and Non-GM-crop Differences Over-estimated](#), the “Genomic Misconception”.

Proposal for a Website of Websites There are simply too many websites (see [ASK-FORCE Organization and Related Websites](#)) and not enough coordination, so

there is a need for networking structures among the most important websites, a *network of networks* with all the fancy new buttons available like RRS, etc. There should be a place where people see with one glance on the first page what news they can expect on various important sites. It should also not be difficult to add possibilities for an individual choice.

Those website connection activities need professional support with some secretarial/managerial help. We must work out ways in which the broad public can easily reach rebuttals on all the myths, facts, and benefits in the debate on green biotechnology. It will not be difficult to establish a platform for a better communication among the most important websites – in the field of agricultural biotechnology, there are a few very successful ones, but this is not the whole task.



The cosine map of 100 words used more than 31 times in the 6,101 documents on Frankenfoods in 2003 (cosine ≥ 0.1)
(From [104])

The task on uniting the most relevant websites and blogs should not be underestimated, see the list already given above [99].

General Views on the Dialogue Related to Regulation of GM Crops

- It should be clear without explanation that each and every rational decision is a combination of facts and values – a decision requires judgment. The agents of judgment are, of course, people, and this leads us to an

As of now, this is just an idea and needs to be discussed with Internet and website specialists. One of the main difficulties will be to establish permanent existence, this is why it would be best to use structures having proofed long years' activities and assured permanence, such as ISAAA, the International Service for the acquisition of Agri-Biotech Applications, www.isaaa.org. After all, the leading webmasters and coordinators agree that it is time to *enhance collaboration through better communication*, see section [ASK-FORCE Organization and Related Websites](#). ASK-FORCE.

entirely different interface – that between scientists and policy-makers.

We should keep this in mind when we concentrate here on the *science* of GM crop regulation. See also the analysis of the debate in [The General Strategic Situation of the Debate About Green Biotechnology Today](#). These philosophical thoughts of Saner are at the basis of the discursive methodology for complex decision-making processes, [121–123]. For details, see below in this contribution in sections [Long Term Discourse and Decision Making Processes](#) and [The Second Generation Systems Approach as a New Decision Making Process](#).

A valid overview on the regulatory science and traceability related to GM crops has been published by Gasson and Burke [124, 125], there is no intention to repeat these reviews.

Biotechnology and Economics

How Economics Are Influencing the GM Crop Debate The example of the Flavr Savr Tomato demonstrates that in earlier times, even in Europe, GM food was well received, but several factors just made it clear that economic success was missing [8, 126–128]. And regulation of this pioneer work needs to get a new look; with modern screening methods, the gene silencing on the molecular level revealed some surprises [129].

Economics play a very important role in the process of technology acceptance: This can be illustrated with the present day feed import situation in Europe. First it should be mentioned, that it is the trade policy of Europe still going the wrong way, which causes a lot of difficulties in the transatlantic dialogue: As Graff et al. [130] explain:

- ▶ European policies blocking genetically engineered crops are conventionally attributed to the concerns of European consumers, but they can be attributed to the self-interests of European industry and farmers as well. Biotech policies maintained in the name of consumer interests are helping European chemical firms to slow their losses in the global crop protection market and are helping European farmers differentiate their conventional crops on environmental and safety grounds, maintain their agricultural subsidies and win new non-tariff trade protections.

The recent development in feed supplies, see Lawrence in *The Guardian* [131], in the EU provides argument, and the reports and letters below give excellent examples:

- Food Chain Dossier 2009: <http://www.botanischergarten.ch/Feed/Food-Feed-Chain-Dossier-20090616.pdf>
- DG AGRI feed report: <http://www.botanischergarten.ch/Feed/EC-DG-AGRI-Rep-feed-situation-UnapprovedGMOs-200709.pdf>
- EU Report on Pipeline: <http://www.botanischergarten.ch/Feed/Stein-EU-Report-GMO-pipeline-LLP-2009.pdf>
- Letter to the President of the EU Commission Barroso: <http://www.botanischergarten.ch/Feed/Letter-big-Producers-Tolerance-Value-Barroso-20090624.pdf>

Strict labeling and thus a discrimination of European meat from animals fed with GM crops will soon be impossible as a political goal due to *economic* reasons – as it is also scientifically not justifiable [132, 133].

An interesting thesis with economic arguments is promoted by Paarlberg [134]: Today, Africa's production of GM crops is exported mainly to other African countries, and this might go on this way in the coming years, so the reasoning that Africans would destroy export opportunities to Europe by developing their own GM crops is not really convincing. But in reality, there is growing concern: Commercial fear over potential loss of export sales to Europe and East Asia is also a reason for mounting pressure on biosafety approvals in developing countries. Consumer misgivings toward GM food in rich countries combined with restrictive import and labeling policies are prompting GM-free agricultural production in developing countries. The long-term costs of these negative trends could be enormous [135]. Good arguments for this view are produced with lots of facts on economics and negative labeling effects of European developed countries, published by Gruère et al. [136–138]:

- ▶ In this context, the marketing decision of avoiding GM ingredients in food items rapidly became a quality attribute employed in the competition among the retail chains of Europe, Japan and South Korea. A report by the international NGO, Greenpeace, which

has encouraged companies to adopt GM-free policies, provides evidence of the widespread adoption of such practices in Europe [139] as follows:

- Fourteen of these retailers have a policy of not selling GM-branded products under their company name for all European countries. These include Carrefour, Auchan, Sainsbury's, Safeway, Marks & Spencer, Coop Switzerland, Coop Italia, Migros, Big Food Group, Somerfield, Morrison's, Kesko, Boots, and Co-op UK.
- Seven of these retailers have a non-GM policy for their own branded products for their main markets (mainly in their home countries). These include Tesco, Rewe, Metro Group, Casino, Edeka, Schwarz group, Tengelmann).
- Out of the top 30 European food and drink producers, 22 have a non-GM commitment in Europe, including Nestle, Unilever, Coca Cola, Diageo, Kraft Foods (Altria), Masterfoods (Mars), Heineken, Barilla, Carlsberg, Dr. Oetker, Arla Foods, InBev (Interbrew), Heinz, Chiquita, Cirio del Monte, Orkla, Ferrero, Northern Foods, Eckes Granini, Bonduelle, Kellogg and McCain.
- Thirteen of these 22 multinationals have a company-wide non-GM policy beyond Europe. These include Diageo, Heineken, Barilla, Carlsberg, Arla Foods, Dr. Oetker, Chiquita, Cirio del Monte, Orkla, Ferrero, Northern Foods, Eckes Granini, and Bonduelle [138].

Some companies even go beyond banning processed products derived from GM ingredients to include requirements on GM-free animal feed in animal products. Virtually all supermarkets sell only poultry fed with non-GM feeds, whereas the policies for dairy products, beef, and pork vary. The usual crude Greenpeace mix of facts and interpretation helped efficiently to push the companies for the European market to go GM crop free [139, 140]. The simple fact of labeling allows opponent NGOs to drive a polemic campaign of pompous "contamination" reports, thus delivering junk science "evidence" that there is some risk involved in the numerous events of minute admixtures of transgenes traces.

In India, there is a clear positive trend visible since some years after some difficulties in the beginning because local traits had to be created for the many

Indian regions and also because there was right from the beginning a black market with illegal cotton traits developing (which often did better commercially than the legal ones. Presently, there are 38 traits of GM cotton in India [141].

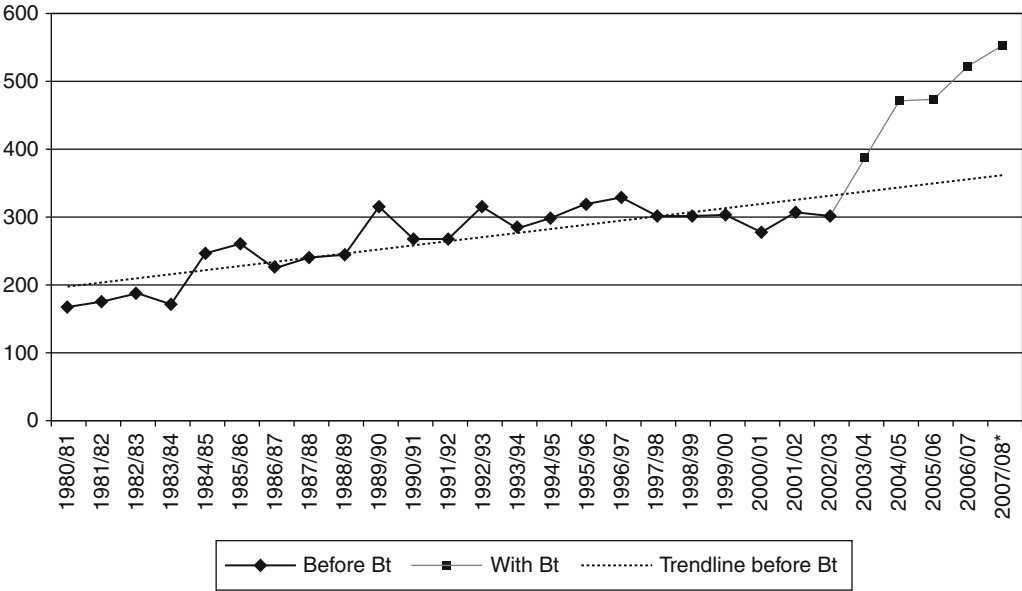
The whole complex story has been recently summarized by [142]:

- On average, Bt-adopting farmers **realize pesticide reductions of roughly 40%, and yield advantages of 30-40%. Profit gains are at a magnitude of US \$60 per acre.** These benefits have been sustainable over time. Farmers' satisfaction is reflected in a high willingness to pay for Bt seeds. Nonetheless, in 2006 Indian state governments decided to establish price caps at levels much lower than what companies had charged before. This intervention has further increased farmers' profits, but the impact on aggregate Bt adoption was relatively small. Price controls might have negative long-term implications, as they can severely hamper private sector incentives to invest in new technology. [142]

At the end of the day the profitability of Bt cotton is now uncontested, see comments of Müller-Jung Frankfurter Allgemeine: [143]

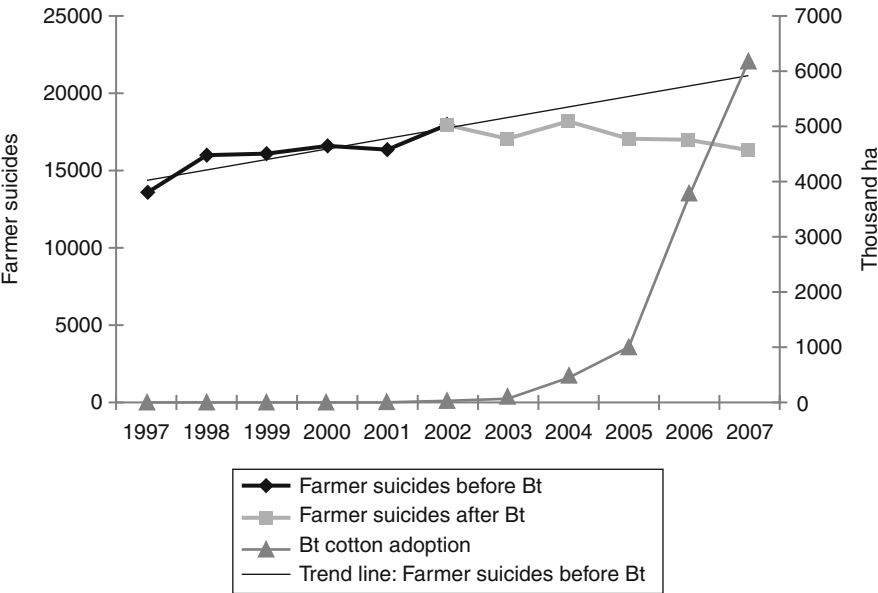
Also the old wrong connection between suicides of Indian farmers and the introduction of GM cotton in India has been thoroughly falsified [144, 145]. This does not hinder activists like Vandana Shiva from continuing with cheap propaganda linking GM crops with the sad tradition of farmers' suicides in India, which started decades before the introduction of GM crops and beginning activities of multinational seed companies. Here are two of the many graphs from [145] (Figs. 6 and 7):

- **Abstract.** Bt cotton is accused of being responsible for an increase of farmer suicides in India. In this article, we provide a comprehensive review of evidence on Bt cotton and farmer suicides. Available data show no evidence of a 'resurgence' of farmer suicides. Moreover, Bt cotton technology has been very effective overall in India. Nevertheless, in specific districts and years, Bt cotton may have indirectly contributed to farmer indebtedness, leading to suicides, but its failure was mainly the result of the context or environment in which it was planted [145].



GM Crop Risk Debate, Science and Socioeconomics. Figure 6

Average cotton yields in India (kg/ha), 1980–2007 (Source: International Cotton Advisory Committee (2008). Note: Data for 2007/2008 is an estimate. From [145])



GM Crop Risk Debate, Science and Socioeconomics. Figure 7

Farmer suicides and Bt cotton area in India, 1997–2007 (Source: Combined data from Table 1 and Table 2. From [145])

► **From the discussions.** The absence of irrigation systems in drought-prone areas (especially in Maharashtra), combined with specialisation in high-cost crops, low market and support prices, and the absence or failure of the credit system, is a clear recipe for failure. It is possible, therefore, that under the conditions in which it was introduced, Bt cotton, an expensive technology that has been poorly explained, often misused and initially available in only a few varieties, might have played a role in the overall indebtedness of certain farmers in some of the suicide-prone areas of these two states, particularly in its initial years. But none of these possible links has been explicitly demonstrated with a sufficiently robust analysis. One implication of this study is the critical need to distinguish the effect of Bt cotton as a technology from the context in which it was introduced. Revealed preferences based on farmer adoption rates and official or unofficial data all point toward the overall success it has had in controlling pest damage and therefore raising average yields in India. In particular, the increasing adoption rate in two suicide-prone states, Andhra Pradesh and Maharashtra, indicates that farmers in these states found this technology economically beneficial.

In contrast, marketing constraints and institutional issues may have played a significant role. Our analysis suggests the need for a better extension system, more controlled seed marketing system, anti-fraud enforcement and better information dissemination among farmers in all regions, before the introduction of any costly new technologies like Bt cotton. Farmers should also be encouraged to diversify their farming and non-farming activities to spread the risks they may incur.

The second implication is that, as farmer suicides are not new or specific to recent cases or to the introduction of Bt cotton, they point toward the failure of the socioeconomic environment and institutional settings in rural dry areas of India. This has nothing to do with cotton or the use of new technology and would suggest many potential policy changes. In several states, such as Karnataka and Andhra Pradesh, some policy changes have already been proposed. Lastly, much more and better federal and state investment could help prevent the 80 percent or more other cases of suicides.

This does not hinder activists like Vandana Shiva from proclaiming Indian farmers' suicides to be the fault of international corporations: [146] and lately also at a Barilla webinar July 20, 2011 in Milano: <http://www.barillacfn.com/en/biotecnologie>, she also does not shy away from connecting the sad tradition of farmers suicides in India with the emergence of GM crops, despite hard facts as demonstrated above. In the same picture you can see her pompous literature list she gives in her curriculum of "over 300 scientific publications in important journals" – a quick test in the comprehensive database of the Web of Knowledge <http://apps.isiknowledge.com/> reveals some 47 papers, most of them in less important journals and magazines – so much about her scientific achievements.

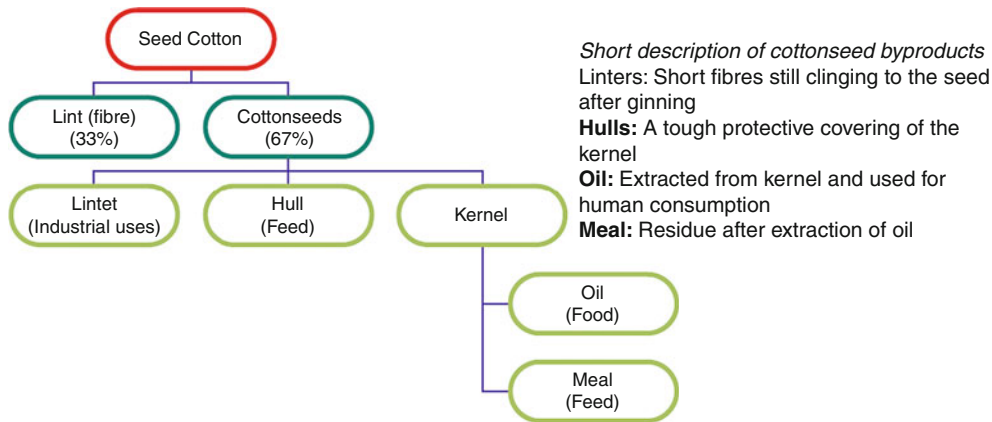
A new perspective is open since 2006 for the production of cotton seed (oil for human consumption), seed meal for feed, made possible thanks to the detoxification (gossypol) successfully done by modern breeding including genetic engineering [147], see the latest summary on the matter (Fig. 8) [148]:

This latest development will open new doors for the cotton production and marketing.

The Political Economy of Biosafety Regulation in Agriculture

An in-depth analysis of how politics is influenced by multiple factors of discursive processes, influenced by economics, has been developed by Graff et al. [149]. They are giving highly differentiated insights in the network of self-interests with some interesting examples of units influencing in their own interest the debate on GM crops: opponents of genetically engineered crops and also industrial units fearing losses in pesticide sales. Often these important socioeconomic elements in the regulatory debate are neglected and it seems to be difficult for all the regulatory analysis to bring together socioeconomic and molecular plant breeding aspects.

► This article develops a political-economy framework to analyze the formation of agricultural biotechnology policies. Going beyond accounts, that largely attribute differences between US and European regulatory environments to consumer attitudes, we consider the impact of what amounts to a Schumpeterian process of "creative destruction" across the entire range of relevant economic sectors and interests. **The analysis**



GM Crop Risk Debate, Science and Socioeconomics. Figure 8
 Cotton seed byproducts (From [148])

suggests that in Europe and in some developing countries a “strange bedfellows” constellation of concentrated economic interests (including incumbent agrochemical manufacturers, certain farm groups, and environmental protest activists) act in rational self-interest to negatively characterize GM technology in the public arena and to seek regulations that block or slow its introduction. In contrast, those interests most likely to experience welfare gains from biotechnology are the more diffused and less informed – including consumers and small farmers. The most profound implications of overregulation of agricultural biotechnology are (1) delays in the global diffusion of proven technologies, resulting in a lower rate of growth in the global food supply and higher food prices, and (2) disincentives for investing in further R&D, resulting in a slowdown in innovation of second generation technologies anticipated to introduce broad consumer and environmental benefits.” [149]

Ayal and Hochman [150], started in some intricate experimental setups working on the cognitive processes underlying choice behavior. With a mix of behavioral actions combined with opinion polls they found that people *do not rely on limited arguments only, but tend to integrate all acquired information into their choice processes*. This could explain the delay in such opinion finding and decision-making processes influencing politics over years, described in the Gartner hype cycles, see [The General Strategic Situation of the Debate About Green Biotechnology Today](#).

Although this would be an epic theme, we shall concentrate here more on the debate of the *Science* of regulation and some discursive elements.

Brazil, A Case Where Politics Positively Influences the Development and Adoption of GM Crops
 Studying the biosafety law of Brazil, the similarities with the European legislation cannot be overlooked: Both legislations are process-oriented and obey strict rules on biosafety assessment, including field experimentation:

A closer look at the Brazilian legislation [151] shows the similarities to the European legislation.

- Article 3. Under this Law, it shall be considered:
 - V – genetically modified organism – GMOs: an organism the genetic material of which – DNA/RNA has been modified by any genetic engineering technique;

And compare some exclusion rules, typically reducing the safety assessments strictly to the process of genetic engineering.

- Article 4. This Law is not applicable when a genetic modification results from the following techniques, provided they do not imply in using a GMO as the receiver or donator:
 - I – mutagenesis;
 - I – the formation and use of animal hybridoma somatic cells;
 - III – cellular fusion, including plant cells protoplasm, which can be produced from traditional culture methods;

IV – the self-cloning of naturally processed non-pathogenic organisms.

The same is the case in the European law: [152], in the introduction the definition of GMOs is given:

- In order to protect human and animal health, food and feed consisting of, containing or produced from genetically modified organisms (hereinafter referred to as genetically modified food and feed) should undergo a safety assessment through a Community procedure before being placed on the market within the Community.

The intention of this “exclusive” definition is clear in this European Law: it should be restricted to GMOs which are wrongly defined as “genetically modified crops,” a scientifically questionable denomination, since in the strict sense of modern genomic science this means to include all crops and horticultural traits having been modified also by conventional breeding. This kind of now false but routine denomination is a symbol for the disregard of proper science in regulation.

A further comparison demonstrates that legislations in Europe and Brazil are both rather strict, the decisive difference is that in Brazil there are clear (political) decision-making rules, whereas these are lacking in Europe. Until lately, the decisions were depending on majority voting rules of the European states, and this caused a lot of confusion and an almost complete stall in decision-making. This is why Commissioner Dalli [153], in July 2010 opened a debate on delegating some important decisions to the national level: Comments from <http://www.gmo-compass.org/eng/news/523.docu.html>

- (13 July 2010) As expected, the EU Commission decided on 13.07.2010 changes in the legal regulation of green biotechnology. Accordingly, Member States should be able to prohibit the cultivation of genetically modified (GM) crops that have been approved EU-wide. As the next step, the EU Parliament and Council of Ministers must agree.

The outcome will again depend on complex negotiations and it is not sure whether Commissioner Dalli and the EU will come to concrete legislative results.

And, except for some modest GMO corn cultivation in Spain, the present day acreage of GM cultivars remains disappointingly low [154].

In contrast to the complex and stalled situation on European GMOs, the case in Brazil documents in the last few years successful regulation of GMOs: Recent reports document steadily growing acres on GMO crops in Brazil: [155, 156]

- The 1st survey on agribiotechnology in Brazil for the 2010/11 growing season showed there was a substantial growth in the adoption rate of biotech soybeans, corn, and cotton. The Brazilian farmers are expected to plant 17.2 million hectares with GM soybean cultivars, or 75.6% of the total harvested surface, in 2010/11.

For a general survey of the Brazilian situation, see the recent publication of Mendonca-Hagler et al. [157], where a clearly optimistic picture is developed. The abstract reads:

- Biotechnology is a Brazilian priority, and has been recognized for its potential to promote sustainable development. The Government recently announced an ambitious program for Science and Technology, which includes strategies to develop modern biotechnology, continuing three decades of public investments on capacity building and infrastructure, aimed principally at the development of technologies applied to health, agriculture and the environment (MCT 2008 <http://www.mct.gov.br/>). Research initiatives have focused on genomics, proteomics, genetically modified organisms (GMOs), gene therapy, stem cells, bio-fuels and nanotechnology, among other biotechnological topics. Research projects in Brazil have been mainly developed in public universities and institutions funded by federal and state agencies, with a minor participation from the private sector [158]. Genomics, an area of considerable success in the country, was launched a decade ago by S. Paulo State Research Foundation (FAPESP), with the organization of a virtual institute, called ONSA, comprising several laboratories with the main task of sequencing the genome of the citrus pathogenic bacterium *Xylella fastidiosa* [159, 160].

The success of this genomic network stimulated biotechnology startup companies and projects with

the focus on other genomes, such as sugarcane and coffee, including functional genomics and proteomics. Following in the footsteps of the ONSA network, the Ministry of Science and Technology created a National Genome Project Consortium involving institutions located in the major regions of the country, with the task of sequencing eight microbial and two plant genomes. Recently, they concluded the sequence of *Chromobacterium violaceum*, a bacterium with exploitable properties, such as the ability to produce a bactericidal purple pigment (violacein) and bioplastics [161]. Later on, several states launched their own genome programs. A group from Rio de Janeiro, part of the Riogene network, recently sequenced the genome of the nitrogen-fixing bacterium *Gluconoacetobacter diazotrophicus*, a sugarcane endophyte involved in enhancing growth of large crops without the addition of nitrogen fertilizer [162, 163], see also the websites of EMBRAPA <http://www.embrapa.br/english> and the Ministerio Biotecnologia e Tecnologia <http://www.mct.gov.br/>.

Agriculture plays an important role in the Brazilian economy, being responsible for ca. 40% of the exports and employing 20% of the active work force. About one third of the Brazilian GDP comes from agribusiness. Traditionally, this country has been competitive in tropical agriculture, supported by strong research programs on conventional and modern technologies. Intense capacity building initiatives resulted in the formation of a critical mass of scientists working in molecular biology and agricultural sciences [158]. Despite these favorable factors, the adoption of GM crops has been delayed due to intense opposition organized by environmental groups and additional difficulties resulting from a conflicting regulatory framework. In this overview, we address the current status of Brazilian biosafety legislation, and discuss the perspectives for the development of molecular biotechnology in Brazil.

This view is confirmed in a recent editorial in *Nature*, [164], interestingly enough with the same emphasis as above on gene sequencing projects which are the basis of independent biotechnological research and development in Brazil.

Also, the latest success of approving regulatory decisions is symptomatic of the positive biotech climate in Brazil: The first fully developed transgenic crop

in Brazil has been approved for commercialization, published in 2007: [165]. The press release of the president of AnBio (National Biosafety Association) Leila Oda emphasizes also the socioeconomic importance of this approval: [166].

Without going into a survey on the Brazilian opponent's activities and reports in detail, here just a typical example published by a medical group (not linked in any way with environmental toxicology) [167] on how science is distorted in order to make a negative and totally unfounded point against glyphosate is given. This paper produces negative toxicological effects on clearly doubtful experimental scenarios: experimental *Xenopus* frog embryos were *injected* with glyphosate, as mentioned in the introduction.

- We show here that sublethal doses are sufficient to induce reproducible malformations in *Xenopus* and chicken embryos treated with a 1/5000 dilution of a GBH formulation (equivalent to 430 μM of glyphosate) or in frog embryos **injected with glyphosate alone (between 8 and 12 μM per injected cell)**. GBH treated or glyphosate injected frog embryos showed very similar phenotypes, including shortening of the trunk, cephalic reduction, microphthalmia, cyclopia, reduction of the neural crest territory at neurula stages, and craniofacial malformations at tadpole stages.

This absurd experiment methodology contradicts all internationally agreed rules on environmental toxicology testing, as described and cited in detail in [168].

But opponents are well organized on an international level, and promptly, the Paganelli paper is cited in many of those reports, here is just one example: [169]. In this extensive report, dozens of papers are cited which do not match the high quality standards of biosafety science; they are cited because they produce negative results related to modern soybean agriculture. The following is an example on how the authors do not even shy away from distorted reporting of published results.

- Very few studies directly examine the effects of GM foods on humans. However, two studies examining possible impacts of GM RR soy on human health found potential problems.

Simulated digestion trials show that GM DNA in GM RR soy can survive passage through the small intestine

and would therefore be available for uptake by the intestinal bacteria or cells [170]. Another study showed that GM DNA from RR soy had transferred to intestinal bacteria before the experiment began and continued to be biologically active [171]. These studies were not followed up. GM proponents often claim that GM DNA in food is broken down and inactivated in the digestive tract. These studies show that this is false.

Actually, if you read the above Newcastle study properly, you notice that the GM DNA is completely decomposed in the colon, the only traces measurable were found in fresh, undigested stomach probes of human ileostomy patients. Reading the summary alone shows the blatant incorrectness of the comments. Two previous studies, after careful reading, reveal the same results [170, 172]. The conclusion therefore is that the interpretation of [169] is false, as confirmed in the latest publication of the Newcastle research team:

- The transgene did not survive the gastro-intestinal tract of human subjects fed GM soya.

A recently published paper of Zhang is seen as a breakthrough in our knowledge on interkingdom relations between plant and animal genomics: [173]. First data, obtained with modern genomic analysis, demonstrate the surprising finding that exogenous plant miRNAs are present in the sera and tissues of various animals and that these exogenous plant miRNAs are primarily acquired orally, through food intake. MIR168a is abundant in rice and is one of the most highly enriched exogenous plant miRNAs in the sera of Chinese subjects. In addition, these findings demonstrate that exogenous plant miRNAs in food can regulate the expression of specific target genes in mammals.

This could lead to erroneous conclusions that horizontal gene transfer is possible also for the antibiotic resistance genes and even for genes expressing Bt toxins into mammals and humans, and one can see already that opponents to genetic engineering take advantage of the news by clear misinterpretation of the results: They use it as an argument for the unforeseen risks of the technology. See the comments of anonymous scientists in GMwatch [174]:

- The study is yet another nail in the coffin of the already discredited 'safety assessment' process for GM foods in

the EU and elsewhere. These assessments do not consider the effects described.

This rather naive statement is typical of the thinking of GM crop opponents: Firstly, they mix up in an unscientific way various categories of transgenes; secondly, they mix up scientific progress and the inevitable adaptation of risk assessment methodology with the present day regulatory rules in place in the laboratories. It is a matter of simple scientific consensus that bio-safety assessment has to adapt in methodology with the progress of genetic engineering: on one side, Zinc Finger and TALES methods (details see [Zinc Finger Targeted Insertion of Transgenes](#) and [TALES: Transformation Method Transcription Activator-like Family of Type III Effectors](#).) with all their precision and elegance are prone to simplified risk assessments after detailed studies. On the other hand, technologies using small RNA molecules will undoubtedly force risk assessment researchers to adapt to appropriate methods of analysis, as already proposed by [175]:

- In the future, the predictive ERA process will need to be flexible and adaptable for analysis of the next generation of crops engineered using RNAi and HD-RNAi. As a first step, regulatory agencies and risk analysts need to become familiar with the science of RNAi and its application to plant biotechnology. A concerted effort is needed to develop a pool of expertise to ask the right questions about potential hazards and exposures, to ensure that relevant data are collected and to characterize uncertainty in risk assessments.

Regulators will have to evaluate the design and implementation of research protocols for laboratory experiments and confined experimental field trials. Scientific questions will need to be answered about off-target effects, non-target effects and the impact of genetic mutations and polymorphisms. Understanding the stability, persistence and half-life of small RNAs in various aquatic and terrestrial ecosystems will be essential for the characterization of exposure pathways. New diagnostic tools will probably be required for the identification and quantification of small RNAs for a range of purposes, including crop identity preservation, monitoring and segregation. Ideally, these tools should have a low detection limit and a high degree of specificity for each RNAi crop, while being relatively inexpensive, functional under field conditions and

operable by individuals with diverse backgrounds and training. With all this in mind, it should be possible for stakeholders, regulators and citizens to develop policies and ERA frameworks for RNAi and HD-RNAi crops. [175]

It is correct that small RNA molecules are considered and used for GM plant improvements, as suggested by [175]. And it is also correct that the risk assessment of GM crops up to now does not specifically include the effects described by Zhang et al., that is, that small miRNAs are obviously passing mammal stomach environments and can be integrated in the organism and even be active genetically. This seems to be routine in the evolution of life (and undoubtedly calls for verification and further studies). And the question arises whether we should *automatically* include in the risk assessment small miRNAs, the answer should be *no*: rather it should be another reason to switch European and UN-Risk assessment to product-oriented mood, following the conclusions drawn in the section on the [GM- and Non-GM-crop Differences Overestimated, the “Genomic Misconception”](#).

The above examples of misleading statements and publications of the opponents lead in a logical way to the following section on the quality of scientific papers:

Peer Review in the Biosafety Science Debate on Regulation

Before we start talking about regulation, a word on the science debate shall precede, which depends on the process of peer review, but it may be flawed in many ways, although there is no real good alternative in sight, despite some attempts to change this situation like the proposal to involve respected science journalists. But there are objections: journalists might become part of the system [176] and give up indirectly their strict impartiality and neutrality – which is, maybe, anyway an illusion. Or it might be that they may simply not have the scientific expertise as demonstrated recently in a contribution of a science journalist in *Nature* [177], extensive critical comments in ASK-FORCE contribution on the Rosi-Marshall publication on aquatic insects, see [178] (more comments about this study are given below). It should also be admitted, that a fresh look of a “greenhorn” might reveal new aspects of the GMO battle.

The quality of biotechnological research is also influenced by the research environment offered to students and is evaluated in a differentiated way for Europe by Reiss et al. [179]. Peer review is a very fragile instrument and needs constant inquiry, as demonstrated also on the Wikipedia website on the subject of peer review http://en.wikipedia.org/wiki/Peer_review. It should also be seriously considered that the present day peer review system is basically “faith based,” as described with convincing details by [180].

A trend toward a magazine style is documented for some important journals as *Nature* and others. The facts show that the percentage of externally peer-reviewed articles has dropped dramatically. Facts will be given in a forthcoming publication of R. Laporte, F. Linkov, and K. Ammann.

We should also include a new element in the reviews and evaluation of science as proposed by Lubchenco [181]: the scientific community should formulate a new *Social Contract for science*.

- This contract would more adequately address the problems of the coming century than does our current scientific enterprise. The contract should be predicated upon the assumptions that scientists will (1) address the most urgent needs of society, in proportion to their importance; (2) communicate their knowledge and understanding widely in order to inform decisions of individuals and institutions; and (3) exercise good judgment, wisdom, and humility. The paper concentrates, according to the zeitgeist of the publication date, too much on environmental issues alone, today we should put into the center of our science strategy debates **humanity as a whole – and this means to take care of the most urgent needs, namely to work on the eradication of hunger.**

However, this process should not be mollified on the costs of hard science. The line between science and pseudoscience is often difficult to draw.

An interesting new aspect has been introduced by the Supercourse Group with Faina Linkov and Ron LaPorte: [182]. It is true that quality control of Internet texts need rethinking, and it is also important to analyze in a critical way peer review of print material: Their comments can be summarized as follows: High-quality, Internet-distributed lectures are not basically different

from written science publications, they also must be documented and references properly. A further element could be a method of quality management introduced originally for the industry by Edwards Deming Wikipedia of Edward Deming http://en.wikipedia.org/wiki/W._Edwards_Deming, who very successfully taught management and quality control also in Japan in the 1950s.

Two more initiatives should be mentioned here, they can be summarized under a kind of *post-publication peer review*.

Faculty of 1,000 System With a total of nearly 84,000 articles reviewed by May 2011, the system has accumulated an important body of comments, see <http://f1000.com/>, the comments, although really critical sentences are not foreseen, the system is now linked to The Scientist and provides helpful orientation about important publications. Some examples have been evaluated by the author [183].

Frontiers of Science Frontiers of Science has been developed over 2 years in consultation with scientists and other faculty, as well as with students and postdoctoral fellows, to address manifest intellectual, logistical, and pedagogical issues, see <http://www.sciencecore.columbia.edu/s2.html> and <http://www.fos-online.org/>

Declaration of a New Global Business Ethos as a Barrier Against Undue Influence on the Publication Policy of Scientific Journals On October 6, 2009, Hans Küng, Josef Wieland and Klaus Leisinger presented the Declaration of a NEW GLOBAL BUSINESS ETHOS at the United Nations in New York http://www.novartisstiftung.org/platform/content/element/3177/Newsletter_3-09_2.pdf.

Although coming from a pharmaceutical company like Novartis, multinational seed companies will (or should) most likely join. Such efforts are important, because there is a constant pressure of undue influence on scientific papers, although resisted successfully by most researchers, but the influence of multinational (in this case pharmaceutical) companies can be hidden but nevertheless powerful:

An example of such influence by units sponsoring scientific journals has popped up in Australia: See the debate around the withdrawal of six Australia-based

Elsevier “fake” journals sponsored by the pharmaceutical industry, see the statement of Elsevier’s CEO Michael Hansen [184] and [185–187]. This kind of influence might still be under control, and peer review is usually functioning in an unbiased way – but the difficulties are deep-rooted, and it is a constant fight for quality, as is summarized comprehensively by Scott [188].

It is a cheap and intellectually intolerable slogan of opponents of genetic engineering in agriculture when they discredit researchers for their relationships with industry, since the great majority of researchers all over the world act as independent persons, although sometimes also funded by industry. The sole quality criteria on science are transparency in applied methods agreed upon by the science community and the reproducibility of the data. For more details see section [More on the Quality of Scientific Publications](#).

In the “dangerous” waters of corporate influence, we need renewed efforts of scientometric analysis, as given earlier in a report of bio-era: [189]. The top part of table 5 reveals the few really successful seed companies in relation to the top universities with agricultural research regarding R&D (Fig. 9):

The calculation rules for the table below:

- The four R&D measures are weighted equally. For example, having 10% of industry patents is just as significant as having 10% of commercialized products. Share of industry R&D output = (share of industry patents + share of industry patent citations + share of industry field trials + share of industry commercialized products)/4 [189].

More on the Quality of Scientific Publications

Coming back to the peer review on the quality of scientific papers, all the above statements do not mean to say goodbye to the factual and methodological scrutiny per se – even after a paper is already published. With a focus on the GM food safety research Chassy and Parrott [168] summarize the criterions on how to judge whether a food study is believable or not: (a) Making sure the samples tested are comparable samples. (b) Testing composition to make sure the tests and controls are comparable. (c) The need for an acceptable balanced and nutritious diet. (d) Why the dose is important. (e) What statistics do and do not tell us.

The top 35 R&D organizations in agricultural biotechnology

RANK	PARENT ORGANIZATION	SHARE OF INDUSTRY R&D OUTPUT
1.	Monsanto	29.82%
2.	Du Pont / Pioneer	10.98%
3.	Bayer / Aventis	10.14%
4.	Dow	5.81%
5.	Syngenta	5.80%
6.	Savia / Seminis	2.57%
7.	USDA	2.38%
8.	BASF	1.71%
9.	Cornell University	1.25%
10.	Stine Seed Farm Inc	1.15%
11.	Florigene	1.08%
12.	University of California	1.05%
13.	Exelixis	0.98%
14.	Iowa State University	0.91%
15.	Rutgers University	0.83%
16.	University of Guelph	0.79%

GM Crop Risk Debate, Science and Socioeconomics.**Figure 9**

Table 5, upper part, with a ranking of biotech companies and universities in the USA, from [189], calculation rules above

(f) The importance of peer review and scientific publication. (g) Guidelines for dealing with conflicting information. (h) Ethical considerations. A very important additional point is emphasized by Kostoff [190]: “Multiple technical experts should average out individual bias and subjectivity.” Two blatant examples of lack of peer review properly done are, among others, discussed in ASK-FORCE (with some additions related to recent publications, all cited in the renewed blog:

- The case of Bt endotoxins supposedly affecting aquatic organisms by Rosi-Marshall et al. [191]

See comments in ASK-FORCE blog No. 3 on Rosi-Marshall et al. 2007b: [178] (including also the latest publications of [192]. The study has been criticized heavily by [193] and [194], the main points of critique, summarized in a letter to the editor of PNAS [195]: No indication about the nature of Bt toxin, nor any data about its origin.

Unscientific extrapolation from lab to field experiments, suppression of an important result of Fig. 3: low toxicity of normal Bt toxin levels for aquatic organisms etc. It is good to know that the authors of the original study admitted some mistakes and tuned down their alarmist interpretation in the first study:

- The case of the Austrian mice experiments supposedly affecting fertility after some generations [196]. After lots of public and scientific debate, which caused serious and unfounded damage to the image of Bt crops, the study results were distributed on hundreds of websites of GM crop opponents. But critique came up, and since there was no publication in a peer reviewed journal available, the rebuttals were not published in journals either. The whole bitter debate is summarized extensively in two ASK-FORCE blogs: [197].

The subsequent official retraction done by the Austrian Government itself is hidden in an European Commission Health and Consumers Directorate-General Summary Record of the Standing Committee on the Food Chain and Animal Health from October 19, 2008: European Commission Health and Consumer Directorate-General, Summary Record of the Standing Committee on the Food Chain and Animal Health Held in Brussels October 19, 2008: http://ec.europa.eu/food/committees/regulatory/scfcah/modif_genet/sum_19102009_en.pdf

See also the published comments of Ammann in [198]:

- Studies that look at non-obvious risks are a welcome addition to the literature, say critics, but poorly conducted studies do more harm than good. “It’s just bad science,” says Ammann. “There are a lot of scientists producing these studies in a very sloppy way. They bolster public fear yet do nothing to resolve conflicts or move the field forward”. And:

But the authors aren’t to blame, says Klaus Ammann, emeritus professor at the University of Bern in Switzerland. They are merely the latest victims of what has become the political gerrymandering of science to bolster and support anti-GM sentiment in Europe. “The Austrian government had exhausted all legal avenues to ban cultivation of GM crops,” Ammann says.

“The Ministry of Health decided to avoid the peer-review process and announce study results at a conference, hide the data from scientists, and let the activists run amok with the help of uncritical media.” Indeed, in the ensuing months the Austrian government has backpedaled. The Ministry of Health responded to a request to interview Zentek or other authors with the following: “We asked the scientists to reevaluate their statistical analysis. Additionally the external evaluation will soon be started. I kindly ask you to wait with your proposal until the reevaluation is completed.” [198]

- The case of a review by Dona and Arvanitoyannis [199]. This review would never pass tests designed by Tang et al. [200], which can detect biased filtering of citations and words: According to Tang et al., it is important to distinguish between *subjectivity classification* retrieved from opinionated and factual statements, and combine it with a multiclass *senti-ment classification*, and to get a better scale by using neutral training examples. An extensive scientific analysis on [199] has been placed in ASK-FORCE with critical comments: [201]

A caveat at the end of this paragraph on peer review is appropriate. Although it is in principle necessary to ask ethical questions, we should first concentrate on the scientific assessment of a professional peer review strictly following a factual agenda such as [168, 202] are demanding. Only then when this filter has been passed successfully, it is important to go into ethical and socioeconomic questions. But as often, it is the farmers and the market regulating efficiently, and – no surprise – they follow quite naturally socioeconomic principles. It is wrong to mix scientific and ethical questions as de Melo et al. and Interman et al. are asking for [203, 204], the result is then to accept for discussion a paper like the one of [205], which has been seriously and repetitiously criticized on a factual basis by EFSA [206–208]. Such papers should not be seen as a publication which takes also into account a “balanced view,” because they are flawed in the first place. Papers from the laboratory of Séralini are then often cited as done by *independent* scientists, which is not very convincing, since digging into the financial support of Séralini and his CRIIGEN lab it is highly interesting to

realize that they also receive funds which come from opponents of GMO technology, such as Sevene Pharma, commercializing homeopathic products which claim to detoxify various toxic products [209] and more. CRIIGEN has been created with the financial support of the retailer Carrefour, which has also contributed financially to certain studies of Séralini and his group. Interestingly enough, Carrefour, the second largest food distributor in the world, sells its own brand of “GMO-free” products. . . Source: [210].

- The result of this discussion: it will be necessary to call for new, Internet-based methods to create a more efficient peer review system. A nucleus of such a system is given in Ron LaPorte’s supercourse system <http://www.pitt.edu/~super1/>.

GM- and Non-GM Crop Differences Overestimated, the “Genomic Misconception”

Early Phase of Risk Assessment In the wake of molecular breeding, in particular with the first successes of “gene splicing,” the safety debates started soon after the discovery of the DNA structure by Watson & Crick [211–213], followed by the Asilomar Conference [214, 215] – see also some historical accounts [7, 216, 217]. The fascination about the novelty of transgenesis was justified, but also overwhelming, and the many unforeseen scientific breakthroughs following were unprecedented in the history of molecular biology. Unfortunately, the enthusiasm also lashed back in an overacting in risk assessment, when the first GM crops went into production. The debate on how GM crops should be regulated started very early with an emerging divide between regulation in the USA and Great Britain, including later the whole of Europe [218, 219]. Some more traces of early disputes about regulatory decisions in the USA and in Great Britain can be seen in letters to Nature in 1992: [220, 221]. Some support tighter regulation including field bio-safety assessments, others fear strangulation of biotechnology research. During the wake of the Cartagena Biosafety Protocol most countries adopted (around 2003) the European way of risk analysis of genetic engineering, emphasizing process-oriented regulation and rejecting product-oriented regulation.

The seemingly absolute novelty of genetic engineering on the molecular level has been contested already in the early days of molecular biology in the 1930s and 1950s with the discovery of cellular systems for genome restructuring discovered with the classic papers of McClintock [222, 223] and with later commentaries of Fedoroff [224, 225], also summarized under “natural genetic engineering” [226, 227].

Molecular Processes Similar in Natural Mutation and Transgenesis

Genetic engineering has been brought into evolutionary perspective of natural mutation by authorities such as Werner Arber: his view remains scientifically uncontested that molecular processes in transgenesis and natural mutation are basically similar [228–232]. In a recent paper, Werner Arber [19] reemphasized those similarities on a broader organismal and evolutionary basis; the abstract reads:

- By comparing strategies of genetic alterations introduced in genetic engineering with spontaneously occurring genetic variation, we have come to conclude that both processes depend on several distinct and specific molecular mechanisms. These mechanisms can be attributed, with regard to their evolutionary impact, to three different strategies of genetic variation. These are local nucleotide sequence changes, intragenomic rearrangement of DNA segments and the acquisition of a foreign DNA segment by horizontal gene transfer. Both the strategies followed in genetic engineering and the amounts of DNA sequences thereby involved are identical to, or at least very comparable with, those involved in natural genetic variation.

Therefore, conjectural risks of genetic engineering must be of the same order as those for natural biological evolution and for conventional breeding methods. These risks are known to be quite low. There is no scientific reason to assume special long-term risks for GM crops.

For future agricultural developments, a road map is designed that can be expected to lead, by a combination of genetic engineering and conventional plant breeding, to crops that can insure food security and eliminate malnutrition and hunger for the entire human population on our planet. Public-private partnerships should be formed with the mission to reach the set goals in the coming decades. “from [19].

The same claim is made with a more organismic view by Hackett [233].

It is therefore no surprise that a natural transgene species has been discovered in a widespread grass genus [234]. An extensive overview on “natural transgenic organisms” is given in the excellent blog of David Tribe GMO pundit on natural transgenics: <http://gmopundit2.blogspot.com/2005/12/collected-links-to-scientific.html>.

Recent publications demonstrate that transgenesis, for example, has less impact on the transcriptome of the wheat grain than traditional breeding [235–237] (more details see [44, 238]).

One should also take into account that many of the conventional breeding methods such as colchicination [239, 240] and radiation mutation breeding [241] can be obviously more damaging to the genome, and it is, in addition, not possible to clearly define what impact the untargeted process could have caused. Or, on the other hand, as [242] have demonstrated, that irradiation-induced wheat – *Aegilops biuncialis* intergenomic translocations will facilitate the successful introgression of drought tolerance and other alien traits into bread wheat. In their review, [243] criticized the biased statements of [244, 245] who focus in an unjustified manner on transgenesis alone when describing unwelcome mutations. Still, it has to be admitted that repair mechanisms on the DNA level are powerful [246–248]. It is thus not logical that opposition within organic farming toward genetic engineering is now expanding also to some of those conventional breeding methods, some go even so far as to reject marker-assisted breeding – symptomatic for the organic agriculture scene, this trend is based on the myth of “intrinsic integrity of the genome” [249, 250], for which term it is not possible to find a proper scientific definition, which inevitably should be based on comparisons [44]. The addition of rejected breeding methods would ultimately lead to an absurd situation where most of the modern time traits would have to be rejected and breeding would be forced to virtually start from scratch.

Basically, many of the first-generation GM crops should be today subject to a professional debate on *deregulation*, and there is good and sturdy reason to state that many of these GM crops should not have been treated in such a special way in the first place, they can

be compared in their risk potential to many crops created with traditional methods.

- This should not be misunderstood as a plea for general deregulation of GM crops, rather for a strictly science-based, risk-based regulation and clearly for a shift from process-based regulation toward product-based regulation.

Dissent over Differences Between GM- and non-GM Crops Causes Transatlantic Regulatory Divide This actually includes a critical questioning about some basic rules of the United Nations Convention on Biological Diversity (CBD). Transgenic crops of the first-generation should not have been *generally* subjected to regulation purely based on the *process* of transgenesis alone; rather it would have been wiser to have a close look at the *products* in each case, as John Maddox already proposed in 1992 in an editorial in Nature [251]. This is also the view of Canadian regulators [252–254], where the *novelty* of the crop is the primary trigger for regulation. This transatlantic contrast has been commented by many [16, 218, 255–258], and although for many years a solution and mediation seemed to be too difficult, contrasts can be overcome:

In a letter to the executives of the Convention on Biological Diversity (CBD), the Public Research and Regulation Initiative (PRRI) http://www.pubresreg.org/index.php?option=com_docman&task=doc_download&gid=490 is asking for a scientific discussion in order to exempt a list of GM crops from the expensive regulatory process for approval, here is only the final statement:

- Bearing in mind that the **method of transformation itself is neutral**, *i.e.*, that there are no risks related to process of transformation, PRRI believes that there are several types of LMOs and traits for which - on the basis of the characteristics of the host plant, the functioning of the inserted genes and experience with the resulting GMO - **it can be concluded that they are as safe as its conventional counterpart with respect to potential effects on the environment, taking also into account human health.**

Unfortunately, there was no substantial reaction from the leading Cartagena organizers.

To be quite explicit once more, this does not mean to exempt transgenesis from biosafety assessment as a whole, but it should say that “several types of LMOs and traits, where the inserted genes demonstrate in large scale commercialization (of course after risk assessment done in due course) can be deemed as safe as conventional counterparts according to several years of beneficial agricultural practice, should be exempt under article 7.4 of the Cartagena Protocol for further expensive and time-consuming risk assessment and regulatory procedures. This motion has now officially been repeated by PRRI (Public Research and Regulation Initiative at the occasion of the COP10-MOP5 negotiations in Nagoya, Japan, see the interventions on the website www.pubresreg.org with recent additions.

In a recent paper, an indiscriminate continuation of food biosafety research is questioned on the basis of all the above arguments by Herman et al. [259] with good reason:

- Compositional studies comparing transgenic crops with non-transgenic crops are almost universally required by governmental regulatory bodies to support the safety assessment of new transgenic crops. Here we discuss the assumptions that led to this requirement and lay out **the theoretical and empirical evidence suggesting that such studies are no more necessary for evaluating the safety of transgenic crops than they are for traditionally bred crops.**

Perspectives for Solutions, a Synthesis of Divergent Views in 2.4

These new perspectives create hope that solutions can be found. Even within the difficult and for GMOs totally negative legal environment of the Cartagena Protocol, there are some slim possibilities:

In a first phase some of the widespread transgenic crops like transgenic maize with the Cry1Ab endotoxin could be exempt from regulation. This is indeed possible according to art. 7.4 in the Cartagena Protocol. In COP-MOP5 2010, in Japan (Fifth meeting of the Conference of the Parties serving as the Meeting of the Parties to the Cartagena Protocol on Biosafety (COP-MOP 5), 11–15. 10. 2010 Nagoya, Japan <http://bch.cbd.int/protocol/meetings/>) it should be possible, to amend

the protocol with the introduction of a dynamics which allows to start the regulatory process with an initial phase focusing on the process of transgenesis, first following procedures proposed for nontarget insects by [260, 261].

Indeed, in COP10-MOP5 in Nagoya October 2010, PRRI www.pubresreg.org has made a request for the exemption of widely adopted Bt maize crops of the endotoxin type of Cry1Ab, see the press release for the context (PRRI press release: http://www.pubresreg.org/index.php?option=com_docman&task=doc_download&gid=586), here the original text as read at the plenary meeting in Nagoya: PRRI Statement on exemptions MOP5: <http://www.ask-force.org/web//PRRI-MOP5/PRRI-MOP5-statement-Strategic-Plan-delivered.pdf>:

- Third, there is an underlying misperception that there are demonstrated cases of adverse effects. This is incorrect. Over the last 15 years GM crops have been planted over a billion hectares by tens of millions of farmers in the developing and developed world. These crops have been grown in numerous different environments, and they have been consumed in billions of meals. The substantial scientific evidence accumulated shows that there are **no** verifiable reports of any adverse effect to environment or human health.

The Strategic plan includes an indicator “Number of reports to the BCH on the identification of LMOs or specific traits that may have adverse effects”. Such an indicator makes little sense, because it is never possible to rule out that any organisms, LMO or non LMO, may have adverse effects. What is crucial is the question whether they are likely or unlikely to have adverse effects, and PRRI proposes that the strategic plan includes these two questions. PRRI is ready to submit examples of categories of LMOs of which the risk assessments and accumulated evidence indicate that they are unlikely to have more adverse effects on biodiversity or human health than their non modified counterparts, and that consequently those LMOs can be exempted from the AIA procedure on basis of article 7.4 of the Protocol.

In future, it should also be possible to shift eventually the focus on the product, making it possible to abbreviate the regulatory process wherever possible and feasible. The ultimate goal of new regulatory

concepts should be to minimize obstacles for new and urgent necessities in crop development, such as Swaminathan and Raven are proposing [262, 263]. The author remains pessimistic, since the whole cumbersome process of legal changes in the Cartagena Protocol is also systematically hindered by a strong anti-GMO lobby, having made its way through the institutions to higher and powerful positions within the Cartagena administration quite successfully, starting from MOP1 all the way up through MOP5, thus influencing negatively all change of regulatory appeasement and lowering regulatory costs. Unfortunately, the recent overview of the European legislation on GM crops does not generate much optimism either: [264].

A second negative trend is triggered by a growing community of risk assessment researchers, who have a vested interest to keep the pot cooking, examples can be downloaded at the website of GENOK www.genok.com and also from the website of the Third World Network <http://www.twinside.org.sg/> with its intricate mixture of activist statements and questionable and peer-reviewed scientific contributions. Other similar examples supporting this view can be downloaded over the Freiburger Oekoinstitut <http://www.oeko.de/> and on the website of ENSSER, European Network of Scientists for Social and Environmental Responsibility <http://www.ensser.org/>

A conceptual framework is proposed by IFPRI/ISNAR in 2002, the International Service for National Agricultural Research [265]; a careful evaluation of process-based versus product-based triggers in regulatory action can also lead to a merger of both seemingly so contrasting concepts into a legalized decision-making process on which trigger should be chosen in a case-by-case strategy:

- Process-based triggers are the rule in almost all countries that have developed national biosafety regulatory systems; there are exceptions, however, where the novelty of the trait determines the extent of regulatory oversight and not the process by which the trait was introduced. While such a product-based approach to defining the object of regulation is truest to the scientific principle that biotechnology is not inherently more risky than other technologies that have a long and accepted history of application in agriculture and

food production, it is less prescriptive than process-based regulatory systems.

Many of the debates on those two concepts suffer from a lack of clear-cut definitions, it will be important to have a close look at the Canadian regulatory system and the definition of PNTs (Plants with Novel Traits). In Canada, the trigger for risk assessment is the *novelty* of the plant rather than the *methods* used to produce it. The difficulties start there, where a clear definition of PNTs is needed to come to a decision. It means that plants produced using recombinant DNA techniques, chemical mutagenesis, cell fusion, cisgenics, or any other in vitro technique leading to a novel trait need to undergo risk assessment in the Canadian system. No wonder the Canadian definition of novel traits is rather wordy, but remains broad minded:

- ▶ A plant variety/genotype possessing characteristics that demonstrate neither familiarity nor substantial equivalence to those present in a distinct, stable population of a cultivated seed in Canada and that have been intentionally selected, created or introduced into a population of that species through a specific genetic change.

Conclusions: There can be no doubt that product-based regulatory approaches are truest to the scientific principle that biotechnology is not inherently more risky than other technologies that have a long and accepted history of application in agriculture and food production, it is also less prescriptive than process-based systems, see for more details McLean et al. [265].

The Costs and Lost Benefits of Overregulation

The Issue

The Cartagena Protocol on Biosafety (CPB) has now been adopted by 157 parties <http://www.cbd.int/biosafety/signinglist.shtml>. It still builds on the principle that GM crop plants might bare risks in contrast to the conventional crops, objective of CPB: <http://www.cbd.int/biosafety/articles.shtml?a=cpb-01>. The huge apparatus on risk assessment based on this protocol is building on the principle that the mechanism of transgenicity is totally artificial and is not found in nature. Modern molecular science insights have proven

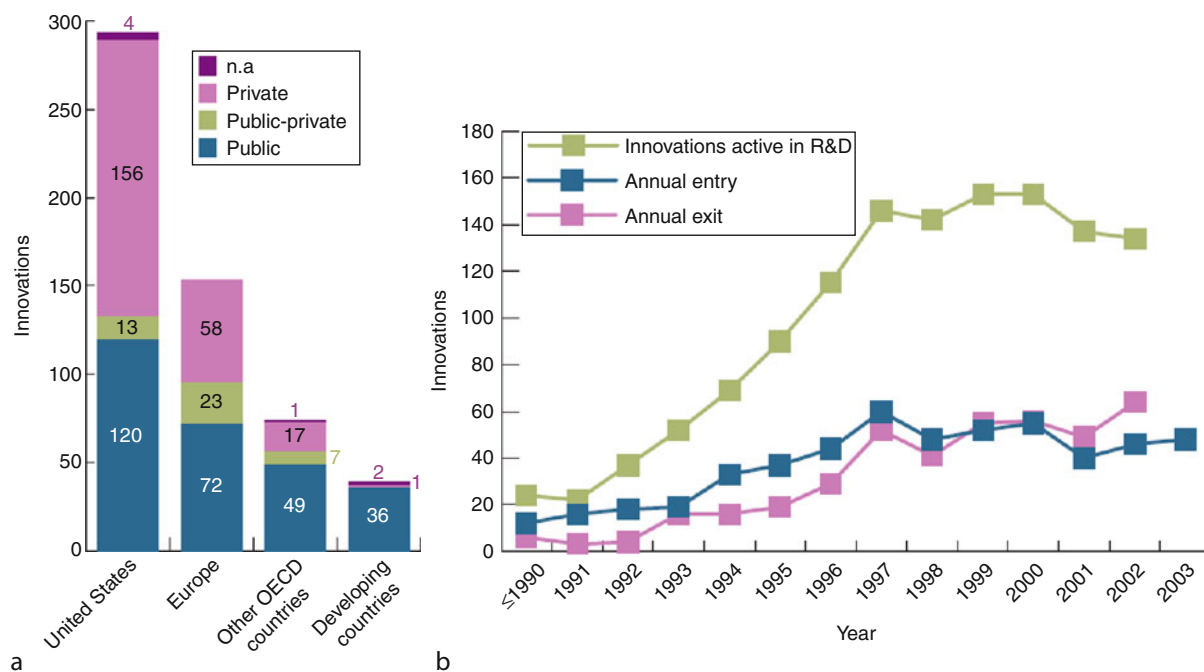
the contrary, as shown in ASK-FORCE AF-9 [201] on the molecular basis of transgenesis. This results in maintaining the concept of an asymmetric risk assessment of innovation of GM crops. The possible exemption of widespread GM crops in Art. 7.4 (Cartagena Protocol on Biosafety, Article 7: <http://www.cbd.int/biosafety/articles.shtml?a=cpb-07>) is not even considered officially up to now.

Summary

An excellent summary graph is given in [266] in Fig. 10b: innovations active in the R&D pipeline were growing at an increasing rate during the period before 1998, but declined after 1998. Apart from competition of reasonably close nontransgenic substitutes, the authors consider one regulatory reason to be the main culprit: The halting of regulatory approvals in 1998 in Europe. Although the authors consider the full extent of reasons still to be conjectural, their data suggest that changes in regulatory environment may have been a cause. In a combination of high costs for lost implementation and high costs for regulatory approvals, the present state and operational experience has grown into a major obstacle of modern crop breeding (Fig. 11).

- ▶ Commentary from Table 1 in [266]: The primary survey combined records from scientific publications, field trial records, and regulatory filings to identify 558 transgenic plants with quality improvements and determine how far they had progressed through stages of R&D by 2004, including those that had only been published in the scientific literature; those that had reached initial field trials (defined as having completed 1–3 field trials), mid-stage field trials (4–9 field trials) or advanced field trials (>10); those that had entered regulatory filings; and those that were commercialized. The secondary survey canvassed expectations of firms and analysts about the likelihood and time frame for future commercialization of transgenic product quality innovations. Complete one-to-one correspondence between individual observations of the two surveys was not possible.

In a recent publication [267] document the same dramatic negative trend for specialty GM crops is demonstrated:



GM Crop Risk Debate, Science and Socioeconomics. Figure 10

Innovation in Ag-Biotech. **(a)** Location and sector of organizations conducting R&D for the 558 transgenic product quality innovations identified. Private sector consists of corporate and privately held firms. Public sector consists of government research laboratories, universities, and nonprofit research institutes. **(b)** Annual entry, exit, and the numbers of innovations active in the R&D pipeline were calculated from observations of the 558 innovations tracked in the primary survey. The number of active innovations stopped growing in 1998, after which those new innovations that entered were more likely to be published and less likely to move toward commercialization. Figure 1 from [266]

Costs and Lost Benefits Worldwide and Europe

An excellent summary graph is given in [266] in Fig. 6 above: innovations active in the R&D pipeline were growing at an increasing rate during the period before 1998, but declined after 1998. Apart from competition of reasonably close nontransgenic substitutes, the authors consider one regulatory reason to be the main culprit: The halting of regulatory approvals in 1998 in Europe. Although the authors consider the full extent of reasons still to be conjectural, their data suggest that changes in regulatory environment may have been a cause.

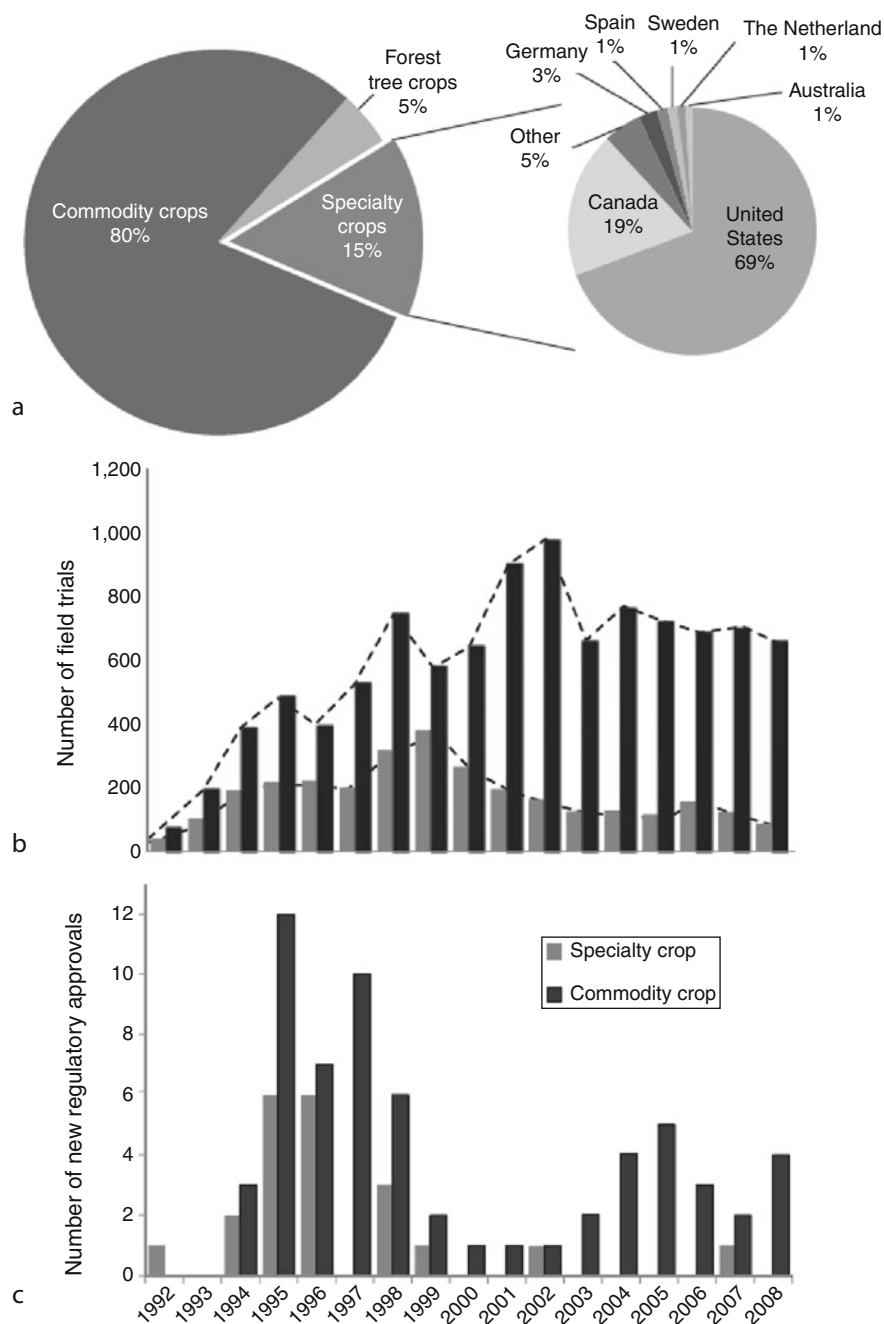
The full extent of the GM crop development pipeline can be evaluated in websites like the Information Systems for Biotechnology alone from the USA, there are (October 23, 2009) 14,204 notifications with 1,586 full field release permits registered in this Database, ISB: Information Systems of Biotechnology: Field Test

Releases in the US: <http://www.isb.vt.edu/cfdocs/fieldtests1.cfm>

Overall, the present day regulatory regime detains public research in molecular breeding considerably due to enormously high regulation costs. More information about this effect on the development of GM trees is in Strauss and McLean [268, 269]; the abstract reads:

- **Against** the Cartagena Protocol and widespread scientific support for a case-by-case approach to regulation, the Convention on Biological Diversity has become a platform for imposing broad restrictions on research and development of all types of transgenic trees.

Some comprehensive tables on the massive costs of regulation of the major commodity crops are given by Kalaitzandonakes [270]. The compliance costs for herbicide tolerant maize alone have been calculated based on the events in 2006 for the USA. They amount to US



GM Crop Risk Debate, Science and Socioeconomics. Figure 11

Field trials and regulatory approvals. **(a)** Using the UNU-MERIT database, field trials conducted in 24 developed countries between 2003 and 2008 were separated on the basis of commodity, forest tree, or specialty crop. From this, the specialty crops were further subdivided based on the country in which the field trial was conducted. **(b)** The numbers of field trial permits acknowledged or issued in the USA are plotted by year for commodity crops and specialty crops. **(c)** The 84 unique transgenic events that have been granted regulatory approval by one or more countries are plotted by year of approval. If the year of approval varied among countries, the first year of regulatory approval granted by any agency for a given event was used. From [267]

\$6,180,000–14,510,000 – a sum most likely to be prohibitive for any trait developed by a public institution.

Another case is reported by Piero Morandini from Italy. A scientific assessment on a field trial on Bt maize is delayed in publication by the Italian Government, although (or because?) it yields very positive results [271, 272].

- The grain yield data (tons/ha, GM crop vs. their conventional counterparts) were rather spectacular: 15.9 vs. 11.1 and 14.1 vs. 11.0, translating into a 43 and 28% yield increases for the P67 and Elgina, respectively. These data have already been released by the INRAN (National Institute for Research on Food and Nutrition, a research institution funded and run by the government) in 2006, albeit without the emphasis they deserved.

The delay in properly communicating these data can be considered as a very costly omission. In fact, taking into account the total area of maize cultivation in Italy together with yield differences, maize prices and pest pressure, **these data translate into a forfeited value of between roughly € 300 million and € 1 billion a year because Italian farmers are not allowed to plant Bt maize.**

A summary of the Lombardia maize case has also been published in Nature Biotechnology [273]. Unfortunately, the original research report is still not published, it is “resting” in an Italian government drawer. . .

The present day regulatory “cropping apartheid” of high tech farming versus organic farming, large-scale farming against smallholders seriously hampers the development of GM crops, which could foster a more ecological production [44, 274] [275] and [276] – in short, Gene Peace instead of Greenpeace.

Costs and Lost Benefits in Developing Countries

Even more drastically, in the developing world, there is regulatory legislation in place hindering the development of transgenic crop breeding for the benefit of the poor, Driessen, Herring, Paarlberg [277–280].

Doubling agricultural research investment per se (no regulatory costs included in the calculation), would reduce poverty in Sub-Saharan Africa by 9% according to Alene & Coulibaly [281]. But these

prospects are seriously hindered and as a result are practically nullified by the exorbitantly high regulatory costs during the implementation phase. Moreover, GM-free private standards set up by food companies and distributors in developed countries have influenced biosafety policymaking in developing countries: Gruère and Sengupta [282] found 29 cases where private importers have affected policy decisions in numerous countries due to irrational fear of export losses. This is based on two generally misleading premises: (1) Europe or Japan represents the only market for exports, and (2) non-GM segregation is too costly. It is amazing to realize, that many of the cases rely on unpublicized lobbying activities, and because of the lack of comprehensive evidence, many cases do not provide straightforward evidence of causality links between importers or traders and policy decisions. There is evidence that development of GM crops in Africa is mainly based on public research, and that the private sector only reluctantly invests in projects for developing countries, although the situation is getting better in the last few years [283, 284].

A blatant case of eco-imperialism is reported from Zambia by Andrew Apel in GMobelus: <http://www.gmobelus.com/news.php?viewStory=234>, where the Norwegian Government has partly sponsored a \$400,000 laboratory, for which GENOK has contributed equipment and training, thus guaranteeing a research policy hostile to GM crops, in accordance with the official policy of the Zambian government, that characterizes GM crops as poisonous. The Norwegian GENOK is a well known anti-biotech NGO, with a very negative attitude toward GM crops, not shying away from spreading myths on allergy caused by pollen of transgenic maize in the Philippines; This is documented in the controversy between GENOK and Rick Roush: <http://www.botanischergarten.ch/Allergy/Traavik-Roush-Philippines-controversy-2004.pdf>, also supported in favor of Genok without a shred of evidence by John Vidal from the Guardian: <http://www.guardian.co.uk/science/2004/feb/27/gm.science>. Typically enough, the laboratory's priority will be to detect and search for genetically modified seeds and crops. Former Zambian researcher Ed. Rybicki, now working in Cape Town, said “that the lab would better serve Zambia and the whole region by looking at genuine threats, studying local biodiversity and even making transgenic

crops themselves”, as reported by SciDev Net http://www.scidev.net/en/news/zambia-s-molecular-biology-lab-fully-functioning-a.html?utm_source=link&utm_medium=rss&utm_campaign=en_news. Indeed, it is rather ironical that many of the biosafety educational efforts undertaken by organizations, highly critical to transgenesis, are turned into the “contrary”: the biotechnological methods introduced in those countries are now also used for research and development of GM crops. A comprehensive report on agricultural biotechnology by Alhassan [285] demonstrates that high regulatory hurdles would hinder a reasonable development of modern agriculture in Africa.

Gruère and Smale [286, 287] report in a carefully calculated assessment that if rice cultures in India, Bangladesh, Indonesia, and the Philippines would be based on present day GM traits, the benefits amount to US\$4,331 million. For the USA, an earlier assessment calculates similar sums of benefits related to the introduction of biotechnology in agriculture [288].

There has been much more written about regulatory costs and their negative follow-ups. Here only a small selection of important papers [130, 261, 289–294] is given.

The Golden Rice Development Hampered Through Overregulation. Biofortification as an Ideal Sustainable Way of Foreign Aid in Agriculture

In the case of the Golden Rice this tedious and costly regulation forced upon the regulatory authorities by the CBD solely based on the process of transgenesis has serious ethical consequences as documented in <http://www.agbioworld.org/biotech-info/topics/goldenrice/index.html> and in [270, 295]. A delay of the introduction of the biofortified rice is directly causing each year hundreds of thousands of children to die or to go blind due to severe vitamin A deficiency. Unreasonable and unscientific regulatory obstacles cause massive delay in approvals, especially in developing countries of S.E. Asia [296–311]. The initiator of the Golden Rice Ingo Potrykus project complains bitterly about the unjustified delays due to overregulation in a Nature article: [312].

Specifically related to the developing world, we should refrain from the old myths that international corporate companies are dominating the field – on the

contrary Public Research is responsible for 85% of crop developments, 7% private local companies, and only 1% multinational companies according to figures from Cohen [284], supported by FAO statistics [313]. The myth that patenting rules are seriously hampering the spread of helpful biotech crops in poor countries has been seriously contested [314–316].

As an example, the Golden Rice project will result in biofortified rice traits, which will be distributed to the farmers free of royalties. The Asian farmers will also be able to multiply seeds without paying royalties. The homepage of the project is the main information source <http://goldenrice.org/>. More about the subject can be found in the important and comprehensive Handbook of Intellectual Property Rights of Krattiger et al. 2007 [317], and more: [318–321].

Biofortification programs are prone to get the highest index numbers in the evaluation system for foreign aid programs of Lempert [322]. Biofortification of indigenous landraces by systematically crossing-in the valuable and royalty free traits to enhance the nutritional value is certainly one of the best ways to sustainably help indigenous people suffering from any kind of malnutrition. In all cases known, the technology transfer is royalty free, secured by contracts.

Use of an indicator to assess the quality and success of developing aid projects defined by [322] reveals that most of the major NGO and UN actors in the field of development are actually providing relief rather than development and are creating dependency by treating symptoms rather than long-term solutions. The indicator points to the specific areas where they need to improve in order to fulfill sustainability criteria including tests of whether aid distorts financial markets and business competition, erodes appropriate government functions, and reverses colonial institutions and ideologies that interfere with sustainable consumption within a resource base.

Estimates in costs for vitamin A capsules are clearly incompatible with the living standard in developing countries; a major distribution campaign would result in millions of dollars. Neidecker-Gonzales [323] produced in their study the following figures:

- Total costs are lowest (roughly US\$0.50 per capsule) in Africa, where wages and incomes are lowest, US\$1 in developing countries in Asia, and US\$1.50 in Latin

America. Overall, this study derives a much higher global estimate of costs of around US\$1 per capsule.

A bibliography of publications of the Golden Rice and Biofortification demonstrates the importance of this field of research; out of a general bibliography of 1,640 references a list of over 200 important papers is assembled: <http://www.botanischergarten.ch/Golden-Rice/Bibliography-Golden-Rice-WOS-KA-20091008-links-abstracts.pdf>.

It should be mentioned that biofortification strategies are also proposed for feed [324]. Straw from harvested crops can be adapted to higher feeding straw quality for cattle.

Conclusions drawn by Ingo Potrykus [325], the creator of the Golden Rice:

- The huge potential of plant biotechnology to produce more, and more nutritive, food for the poor will be lost, if GMO-regulation is not changed from being driven by “extreme precaution” principles to being driven by “science-based” principles.

Changing societal attitudes, including the regulatory processes involved, is extremely important if we are to save biotechnology, in its broadest applications, for the poor, so that public institutions in developing as well as industrialized countries, can *harness its power for good*.

As a whole, the new, well-documented review paper of Adenle [326] delivers overwhelming evidence that GM crops are urgently needed in the developing world:

- The world needs fast and reliable solutions to fast growing population and the problems of hunger, malnutrition, ravaging diseases, poverty and global warming crisis. One of ideal technological innovations such as GM technology can be part of solutions to these problems. It is imperative to understand that GM technology cannot establish its ground if continuously faced with the baggage of constraints as discussed above. Moreover, it is not surprising to gather from a variety of literatures that most developing countries lack capacity building and still struggling with the establishment of biosafety system that can facilitate GM field trials and commercial release of GM products. Some of the challenges associated with the development of modern biotechnology still boil down to the fact that individual country government and international organisations have not clearly identified

a coherent strategy and enabling policy instrument to deal with the problems. While some progress have been made on GM technology in terms of research and development, capacity building, and biosafety regulation in developed countries and a few developing countries, concerted effort is still needed to make it an accessible technology for every country. [326]

The Dispute Between Scientists and Opponents Today

The Role of Some Activist NGOs in the Debate

There is a continuous need for dialogue with regulators, the public, and specifically consumers, since the new technology emerging from modern life science is affecting all aspects of human life, including food, reproduction, etc. We do have an unfortunate trend toward irrational and antiscience argumentation in the GM crop dispute as clearly diagnosed by [327] in his book “The March of Unreason”, see also [328, 329].

This said, we should not create misunderstandings. There is no room for appeasement politics today when it comes to the activist NGOs like Greenpeace and Friends of the Earth, or websites like the Institute of Science in Society (I-SIS) and GM-Watch. Those professional organizations have proven repetitiously not to be interested in peer-reviewed science in a debate on the science and the sociocultural issues. They rather rely on unconfirmed reports in order to follow their own ideological and commercial interests. Any rational discourse with such organizations would be very welcome, but needs to be based on the latest peer-reviewed science. Their usual tactics is to appeal on fear. A good example from Greenpeace has been described on the EFB forum website <http://www.efb-central.org/index.php/forums/viewthread/13/> about baseless accusations that 1,600 sheep have died from feeding Bt cotton leaves. A critique on the distorted picture on Indian cotton cultivation by NGOs is given by Herring [330] with lots of figures, facts, and extensive documentation.

Another blatant example of junk science has been launched recently by Greenpeace on You Tube “Genetic engineering: The world’s greatest scam?” <http://www.youtube.com/watch?v=1H9WZGKQeYg> full of misinformation and hatred against multinational seed companies.

We are also confronted with violence – activities clearly documented and justifiably named and pursued as terrorism [331]. Also, in Europe, there are regularly occurring field destructions [332], which hamper seriously biosafety research – what an irony! Eco-terrorism is not confined to Europe, problems of such kind are very real also in the USA [333]:

According to the Federal Bureau of Investigation (FBI), the Earth Liberation Front, together with its sister organization, the Animal Liberation Front (ALF) has committed from 1997 to 2003 more than 600 criminal acts that have resulted in more than \$43 million in damages. Moreover, attacks have been perpetrated in virtually every region of the USA against a wide variety of targets.

Recently, Greenpeace destroyed government field research in Australia [334] and defended the act of eco-terrorism with very thin arguments – and promptly lost lots of supporters and sympathy: Even some old friends and supporters of Greenpeace (but not all) distanced themselves from the action: [335]. A list of field destruction actions in Europe has been compiled by Marcel Kuntz [336]. This list, far from being complete, demonstrates that activists have lost their moral compass in recent years: [337, 338].

One of the best rebuttals of cheap anti-GMO propaganda coming in attractive book editions, widely distributed in international events by the author Jeffrey Smith [339, 340] has been published on the Internet by Bruce Chassy <http://academicsreview.org/reviewed-content/genetic-roulette/>. It is actually a scientific comment, section-by-section, based on the best available peer-reviewed literature.

More chagrin emerges from the mounting pressure from within the academia, where, for instance, German university leaders in Giessen ordered to cease field research on GM crops which is unwelcome in the eyes of the extremists, [341] and there are serious complaints about the difficult atmosphere for biotech researchers in Germany [4].

Another symptomatic row is presently taking place in India, related to the approval of Bt brinjal, where activists are in a desperate attempt to stop the regulatory approval of Bt brinjal with outrageous and completely unfounded rumors like “GM brinjal will render the soil sterile,” But contradictions have been posted as well: the most recent and comprehensive

summary report published by Kameswara Rao [342], which is a review of massive evidence for the safety of Bt Brinjal and the detrimental heavy use of pesticides for the production of conventional Brinjal. It is ironic that one of the main arguments for proponents of the Bt Brinjal moratorium in India is now seriously questioned. There was the seemingly clear evidence on a crop biodiversity center for Brinjal in India, which called for extra protection of indigenous genomes. But recent extensive genomic analysis has clearly demonstrated that Brinjal is originating in Africa [343].

As an exemplary dispute, you can also follow the exchange of letters between the Public Research and Regulation Initiative (PRRI) and Friends of the Earth (FoE) [344]. Some of those anti-GMO activist groups get hefty funding from governments in the EU, as documented accurately by Andrew Apel and his GMobelus website: Europe’s massive funding of world-wide activism. Compare also his recent article on the same subject, focusing on global aspects: [345].

The current set of arguments of GM crop opponents is often a mix of anti-American, antiglobal, post-modern, and even antiscience notions, [346], a strategy which has now been taken over very successfully by NGOs like Greenpeace and Friends of the Earth as global actors. These leading protest forces have helped, particularly in Europe, to build up a postmodern negative picture of biotechnology as a whole [347]. In this light, it is easy to act as “opinion leaders” with pseudoscientific arguments. The feedback mechanisms through the media and a network of citations of all the flawed stories make it possible for the global opponents to maintain confirmation of negation mechanisms. We are in a situation where the opponents already try to claim victory, penetrate highest political levels in governments and international organizations like the United Nations, some produce strikingly flawed reports on GM crops.

An analytical article about media and NGO activities in New Zealand has been published by Motion and Weaver [348]: by attracting media attention through dramatic protests, Greenpeace risks to jeopardize its reputation. The abstract reads:

- The challenges of attracting positive media attention are likened to a contest in which various organizations

attempt to promote and circulate their version of events; however, this is particularly difficult when attempting to circulate less established, unpopular or critical knowledge. Although complying with, and managing, news values is an important starting point, the need to move beyond news values to consider the commercial values and realities of media organizations is highlighted. In this paper, a case study is undertaken of the Greenpeace media relations in New Zealand when a proposed controversial expiry of a moratorium to release genetically modified organisms into the environment. The predicament for Greenpeace is that in attracting media attention through dramatic protests it risks jeopardizing its reputation as a credible news source that can influence the framing of news stories. Insights are offered into the need for organizations to understand and manage the story or knowledge to be circulated and comply with contradictory news values.

Related to this paragraph on NGOs, it is necessary to write a word on the press: Newspapers and other media usually are mirroring what is important in the public debate, and the NGOs are clever in manipulating both the public and the press, after all, it is easy to provoke with fear and scaremongering, and the majority of journalists of all calibers are also committed to their own product, position, and its commercial situation.

A classic example is the coming and going of the Frankenfood Myth, see Fig. 3 and http://en.wikipedia.org/wiki/The_Frankenfood_Myth. Interestingly enough, this myth had its sharp peak in the press statistics around 1998 (see Fig. 3) and since then it has vanished from the headlines [104] as a major buzz word.

Those mechanisms have been precisely described by Burke for the situation in Great Britain some years ago [349]. But it is also clear that in the last 5 years more balanced voices appeared in the press, although there is no room to extend this topic here, just one recent example from the London Financial Times may suffice [350].

The GM Crop Battle, the Dispute Among Scientists, the Use of Strong Language

First, let us not forget some words of Antony Shelton [291], the most important words can translate into

a slogan: “Quality of science must back up personal opinions,” the abstract reads:

- In agricultural biotechnology there are roles and responsibilities of scientists, scientific journals, the public media, public agencies, and those who oppose or advocate a specific technology and serious consequences for science in general when those roles and responsibilities go awry. Scientists may feel the pressure of competition, especially in an academic setting. Personal views may continue to decide which issues one will work on, but the quality of science must back up those personal opinions. Common sense tells us that scientific inquiry and the publication and reporting of results to the scientific community and general population should be performed with high standards of ethical behavior, regardless of one’s personal perspective on agricultural biotechnology.

One of the arising problems is that there has been recently a tendency to mollify peer review for the sake of politically correct so-called critical views of genetic modification of crops, with some blatant examples of flawed pseudocritical papers having passed for publication in highly respected scientific journals – a few examples have been commented by [351]. Some of those papers just passed due to flawed peer review, others passed despite rejection by some peer scientists, obviously for the sake of public debate (and for the promotion of the journal), see as an example the rather thin justifications of the editor in chief of Lancet Richard Horton to go ahead with the publication of Pusztai’s rat experiments [352–356]. For more details about this controversy, see in ASK-FORCE on Pusztai [357], it is an anatomy of the case in 46 pages on the Pusztai affair, which had a big influence on the regulatory climate on GM crops in Great Britain and the world.

It is only between 2005 and 2011 that a certain fatigue of new negative arguments against GM crops is developing, and it is interesting to note that opponents, lacking real negative health and environmental effects, now shift their emphasis on negative arguments in socioeconomics. There are hardly any new issues in food safety and environmental impact to be dealt with in the last few years. This might also be the reason why in a desperate routine of repetitious “negative,” GM crop stories get into journals, often also on rehearsed

events which have been clearly rebutted scientifically many years before. Those “news stories” often pass uncontested and get printed in “news” media due to a mix of short memory effects of uninformed editors and readers of all kind, or worse, they are purposefully repeated by activists counting on short memory of press and public.

A strange effect should also be mentioned that scientists, who defend good science in biosafety research, sometimes get blamed because they use straightforward language when criticizing flawed papers. A paper on such debates has been published by Nature [177], see the comments in a contribution of ASK-FORCE [178] on a paper on aquatic organisms supposedly harmed by Bt toxins of GM maize by Rosi-Marshall [191] and [192]. There are several controversial hints in this Nature story put forward by science journalist Emily Waltz, who is neither specialized nor experienced in the hot scientific regulatory debate on GM crops, suggesting that to criticize flawed papers with “strong language” is detrimental to the progress of scientific research. This statement was supported by interviewed writers such as Ignazio Chapela (famous for starting the controversy of the Mexican gene flow of transgenic maize with a letter to Nature [358], which later turned out containing insufficient evidence for publication [359], see the latest summary in [360]. Another interview Waltz conducted in the cited *Nature* piece with David Schubert, who tries as a pharmacist to give advice in biosafety rules of GM food, and with his strong anti-corporate mood publishes fraud accusations against pro-GMO scientists [361]. Both interviewees Chapela and Schubert defend independent scientific whistle blowing, but themselves they have a proven negative agenda about GM crops, see more controversy papers: [295, 362, 363]. In the meanwhile, several letters to the editor of Nature have been written commenting the feature of Emily Waltz in Nature, they are all cited in [178], the majority is not supporting her thesis.

Incidentally: Strong language has been used before in the history of science, remember some really bitter and hefty disputes about the history of discovery of the double helix structure of DNA between Watson and Crick [216], who later made their peace again.

Other numerous examples of a fight out in the open are documented about evolution when Darwin

proposed his revolutionary ideas. Two citations of strong language may suffice: in a debate on natural selection [364] writes on a dispute with William Bateson:

- By these admission almost the last shred of that teleological fustian with which Victorian philosophy loved to clothe the theory of evolution is destroyed. Those who would proclaim that whatever is right will be wise henceforth to base this faith frankly on the impregnable rock of superstition and to abstain from direct appeals to natural fact.

Another clear example of sharp and relentless scientific controversy on evolutionary biology with strong language has been described in detail by Strick [365], among the numerous juicy examples:

- His [Bastian's] tone was sharp in response to Huxley's public accusations that his technique was sloppy (a much more high-powered attack than Huxley ever adopted in private when attempting to correct young scientists). Huxley replied with an equally sharp tone, now saying sweepingly that “what Bastian got out of his tubes was exactly what he put into them,” *i.e.* contaminants.

And one last word about strong language: The word “abuse” has been printed by Nature in the Battlefield paper [177] very prominently in the subtitle, when attacking a group of authors including me who criticize flawed papers in the GM crop debate with blunt, but still polite words – what an irony! – And to be quite clear, no complaints from my side. . . .

Negative Effects of Modern Breeding Methods in Food and Environmental Safety do (or Should) not Pass Strict Scientific Procedure Rules and Peer Review or They Are Based on an Unscientific Focusing on Transgenesis Instead on Management Mistakes

If researchers would follow strict procedural rules, the world of scientific biosafety debate would be far less complex, here are a few papers standing for such in fact uncontestable rules: [168, 260, 261, 267, 312, 366–369]. It is a fact that for some years basically no new arguments against agricultural biotechnology (in particular clearly related to transgenesis) on an agronomic base

can be put forward for the most widespread crops, which have run through multiple regulatory processes in many countries.

This does not mean that transgenic crops are completely free of problems, but, in fact, it is that in comparison with conventional crop problems these are minor and manageable in a more efficient way. One of the basic mistakes of GM crop criticism is the unilateral focus on the risks of transgenes inserted, instead of comparing, in a fair and scientific (holistic!) way, with conventional cropping [370].

Still, a growing number of herbicide tolerant weeds are emerging: [371–374]. Powels [375] rightly points to the monotonous fields of glyphosate-resistant soybean landscapes, where the herbicide-tolerant weeds emerge more rapidly:

- Indeed, in spite of longterm use, the evolution of glyphosate-resistant weed populations in non-GRC, burndown systems has been very limited. Thus, functionally competent gene traits endowing glyphosate resistance are relatively rare and not easily enriched in plant populations [376], [377]. This is why glyphosate is a remarkably robust herbicide from a resistance avoidance viewpoint. However, as reviewed above, it is clear that, where there is very intense glyphosate selection without diversity, glyphosate resistant weed populations will evolve. In particular, the evolution of glyphosate-resistant weed populations is a looming threat in areas where transgenic glyphosate-resistant crops dominate the landscape and in which glyphosate selection is intense and without diversity. [375]

But it is also a fact that the emergence of glyphosate-resistant weeds is happening on a much slower pace than that of conventional herbicides [378].

Some critical science journalists question the strategies and behavior of the global opposition players. In a kind of last bid, questionable reviews are published, either containing lots of negative *assumptions* [379] or wrong toxico-analytical concepts resulting in an exaggerated risk assessment for nontarget insects as the lacewings as promoted by Hilbeck et al. [380–382] and contradicted clearly in Romeis [383]. Other examples of questionable eco-toxicological conclusions have been drawn by producing or reviewing flawed data or statistics, or drawing questionable conclusions, see the debate on Ermakova's flawed rat experiments: [384],

more details in a contribution to the ASK-FORCE [385]. Typical other examples recognizable on filtered citation lists are Dona et al. and Séralini et al. [199, 386]. Séralini conducted his experiments in disrespect of the internationally approved rules of biosafety experiments established by the OECD [387, 388] and also avoided the citation of certain contradicting peer-reviewed references. Many of those papers have been or will be treated in ASK-FORCE [389], where you can read about new or recently updated ASK-FORCE contributions, for more details see section [ASK-FORCE Organization and Related Websites](#).

It also must be said (remember Saner's statements at the beginning of this section) that vested interests can be spotted with some biosafety researchers, who are in need of research grants and thus paint a negative picture on biosafety; they symptomatically have difficulties to distinguish between the “nice-to knows” and the “need-to knows.” Example: see the ASK-FORCE contribution [178] on the publication of [13], a paper which is flawed in several ways. It has been completely rebutted by Shelton et al. [14], the questions asked in the Lovei paper are irrelevant for Bt maize cultivation, since the Bt-toxin-technology is overwhelmingly beneficial for majority of nontarget insects [390–394]. One of the major flaws of the Lovei paper is that they used low quality prey for their laboratory feeding studies. A thorough analysis of risk assessment research has been recently published by Raybould [261]: We need to carefully distinguish between basic ecological research and purposeful and targeted risk assessment research which concentrates on the real agronomic risks and needs [395, 396].

The question and negative answer given in the letter of the Public Research and Regulation Initiative (PRRI) to the Secretariat of CBD [397] is fully justified, *and PRRI stands ready to expand on the points made in this letter*.

- 1. Are there LMOs or traits that have caused adverse effects?

No. Since the first application of genetic modification in the 80s, many thousands of field trials have been conducted with GM organisms (to date mostly plants), and since 1996 many hundreds of millions of hectares have been planted with GM crops by many millions of farmers and consumed by hundreds of millions of

consumers in developed and developing countries, without any verifiable reports of adverse effects on the environment or human or animal health.

In fact, taking a broader look, experience with those GM crops has shown environmental and socio-economic benefits in terms of increases in yield, significant reductions in use of pesticides, fossil fuels and soil erosion, less mycotoxins in grains, as well as increased farmers health and income.

Final remarks: Coming back to the first statement of Saner [120] given under [General Views on the Dialogue Related to Regulation of GM Crops](#), value-laden scientific activity cannot be avoided, but minimized – if you refrain to work with flawed data, with filtered citation lists, and with reviews pontificating on negative assumptions. The only remedy is to work with high-quality data produced in a methodologically transparent way following international agreement.

It is appropriate to end this rather pessimistic section with a positive note, not free of irony: As Gupta [398] recently stated, there is hope that the introduction of strict biosafety rules in the Cartagena Protocol, originally aiming at a slowing down or even at stopping the transboundary movement (and indirectly development) of GM crops, now seems to turn into its contrary:

- Through analyzing the dynamics of GMO-related information disclosure to the global Biosafety Clearing House (BCH), I argue that the originally intended normative and procedural aims of disclosure in this case to facilitate a GMO-importing country's right to know and right to choose prior to trade in GMOs are not yet being realized, partly because the burden of BCH disclosure currently rests, ironically, on importing countries. As a result, BCH disclosure may even have market-facilitating rather than originally intended market-regulating effects with regard to GMO trade, turning on its head the intended aims of governance by disclosure.

Debate Improvements: What can we do to Enhance the Situation?

Foremost, it is important to *shift from pro-reactive to proactive mode*. This does not automatically mean to filter away negative views on GM crops and to organize

a eulogy on the benefits, the pro-active mode should actually engage a new mode of debate, which is more discursive, more structured and definitely concentrates on a solution-oriented decision-making process. It is time for action – as far as a strict scientific view is allowing this. There are several websites working hard on sorting out the strictly science-oriented messages in biotechnology, as mentioned below. We should not, as it often happens, in our struggle against the negative pseudo facts focus on the risk alone and thus trap ourselves in a negativistic perspective.

Rather we should address in a balanced way the obvious (or lost) benefits as well. But this alone will not provoke a turnaround. This shift must be embedded in a discourse with concerned people and organizations and it must clearly oppose untruthful strategies of the global protest corporations and thus also refrain from using the same countertactics. One of the appropriate organizations for this activity will be the two platforms: (1) Public Research and Regulation Initiative PRRI www.pubresreg.org run by public researchers and (2) also the European Federation of Biotechnology <http://www.efb-central.org/>, so that public science will get a more important place in the international regulatory debate (but also where private seed companies are not fundamentally battled in a naïve neo-Marxist scheme). In many meetings strictly based on science and organized by PRRI, both platforms are well received. The project outline can be described as follows:

ASK-FORCE Organization and Related Websites

There is a flood of papers which cast doubt on the GM crops already regulated in many countries. Most (if not all) of these papers are written in a bad quality, either with flawed methodologies not internationally agreed upon, or with conclusions which are not supported by the data [13], rebutted by [14], details see in [178]. There are also many reviews published in a scientific style, but unfortunately either with a strongly biased set of references or with unsupported assumptions and doubtful conclusions – contradicted by peer-reviewed publications often not cited. In some cases, the flaws are more hidden: Experimental data are achieved on clearly theoretical schemes, working with outdated Bt maize and nontarget butterflies which have in their

biology, in nature, no connection to maize fields: [399]. It is therefore important to set the record straight and to try to rebut at least the most important and blatant cases.

Within an EU project with Marc van Montagu and Piet van der Meer, which has been granted to PRRI, a blog was launched with the name ASK-FORCE on the PRRI website www.pubresreg.org with the secretarial help of Kim Meulenbroeks (until 2008) and presently Zuzana Kulikova. A list of about 130 items [400] has been compiled with international help and will be entered step-by-step in the grid of the following six sections. (1) General (2) Human and Animal Health (3) Environmental Safety (4) Agriculture (5) Public Perception (6) Developing Countries.

Up to now, 11 contributions have been published on the Internet; for links and contributions see [389]. These were reviewed by the experts of the steering committees of Public Research and Regulation Initiative and the European Federation of Biotechnology, some also by the experts united in the blog community of AgBioWorld <http://www.agbioworld.org/>. All three lists contain some of the best specialists on green biotechnology from all around the world for reviewing and commenting.

In order to become more proactive, we need to develop forward-looking strategies. It is up to the scientists to ask questions to the opposition, and in particular to the professional distorters of the scientific facts. This must escalate into public campaigns if (what is to be expected) those specific questions are ignored. Carefully built contacts with science writers are important here, as a help for networkers a selected list is given here <http://www.ask-force.org/web/ASK-FORCE-Summary/Contacts-ASK-FORCE-2011.pdf>

Long-Term Discourse and Decision-Making Processes

Let me first be quite clear that I think a dialogue with the professional protest corporations is, as a rule, a waste of time (specifically Greenpeace and Friends of the Earth, not to mention some other organizations). Their only interest is to keep the pot cooking and make sure that the population remains in a state of fear. They should be addressed with a confrontational strategy, which is included in ASK-FORCE. Often such

NGOs get the willful help of the press, which acts according to the old proverb (Macbeth, Shakespeare) “evil always fascinates – goodness rarely entertains” [401], see also the arguments produced by Andrew Moore [402]. While some press products concentrate on mirroring public concerns, a press more or less close to boulevard strives to foster its marketing with the help of sensational headlines, creating stories which sell better, but indirectly they are exacerbating the problems. We are also not going to talk about a special discourse, as described by Erjavec [403], related to the politics of the EU commission.

Nevertheless we have to address all segments of the public with its concerns, feelings, and interests. And the discourse we are going to concentrate on is solution oriented. This should be done according to the discursive rules of the management strategies of the second generation, the *Systems Approach* (see under [The Second Generation Systems Approach as a New Decision Making Process](#)). As a basic reference with description and citations, see the classic book of Churchman [79]. If we follow some ground rules, this should not be too complicated.

The Second-Generation Systems Approach as a New Decision-Making Process

Instead of making questionable concessions (example: “let’s not talk about transgenic crops” as often done by Nestlé and Unilever, with notable exceptions [404] within these two companies!), the dialogue should be organized in an atmosphere of “Active Listening” [405] and understanding in which, apart from the strict rules of scientific argumentation we should send signals that the new technologies also trigger socioeconomic and cultural feedbacks. This will be the key to solve *Wicked Problems* [406], which contain also sociocultural elements besides a set of hard, often contradictory facts [122]. In his usual cynic precision, George Bernard Shaw defined the ultimate problem in the dialogue between scientists and lay people: “Every profession is a conspiracy against the laity.”

The new discourse is not about the usual stakeholder meetings; rather it is about instigating modern planning processes of the second generation in evidence based but open ended decision-making processes. This *Systems Approach of the second generation*

contrasts to linear planning with predetermined targets and dominating deontic thinking (e.g., of the industrial corporations and government agencies), it contrasts also to *the Systems Approach of the first generation* (e.g., Apollo moon landing with clear target).

The Rationale of New Management and Decision-Making Processes

- Some problems are so complex that you have to be highly intelligent and well informed just to be undecided about them. Laurence J. Peter [407]

These new strategies should dissolve the traditional stakeholder concept in favor of a much more efficient system respecting *different kinds of knowledge* and other rules (such knowledge differentiation is also known from learning processes, which are related to our decision-making dynamics [408]).

There are more practical reasons to employ into the Systems Approach and its concept of different kinds of knowledge, as Zwart [409] rightly emphasizes: Ever since we have realized that the low number of human genes (approximately, 22,500) cannot be interpreted as a narcissistic offence, since organisms are so highly complex, including the emerging consciousness of our human brain, genomics takes us now beyond a genetic deterministic understanding of life, this must have consequences on societal research and debate as well. Policies for self-improvement will increasingly rely on the use of complex interpretation. *Therefore, the emphasis in our discourse must shift from issues such as genetic manipulation and human enhancement to issues involved in governance of novel forms of information.* The same can be said on the side of agriculture. Ikerd [410] develops with the means of the systems approach a more holistic picture of agricultural management.

Fairclough [411] as a linguist gives an in-depth and critical analysis on discourse related to globalization with lots of facets, and again with a totally different set of terminology, he also presents negative examples of discourse. Objectivism treats globalization as simply objective fact, which discourse may either illuminate or obscure, represent or misrepresent. In the Churchman systems approach, there is no such thing as an objective approach, rather it is objectivation. Ideologism focuses upon how particular discourses of globalization

systematically contribute to the legitimization of a particular global order which incorporates asymmetrical relations of power such as those between and within countries.

Scoones et al. [412] come to similar conclusions as the Churchman school, but this time related to agricultural policy, the paper explores the national and transnational character of mobilization against GM crops in India, South Africa, and Brazil in the 10-year period up to 2005. The paper argues for a better understanding of national political and economic contexts which must be taken into account, alongside on how the GM debates articulate with other foci for activism and the complex and often fragile nature of alliances that make up activist networks. It is important to understand that the debate about GM crops has become a much wider one: about the future of agriculture and small-scale farmers, about corporate control and property rights, and about the rules of global trade, see also the new report of the Royal Society [18]. In sum, a debate should not just focus on the pros and cons of a particular set of technologies – after all, they have proven safe – it is more about politics and values and the future of agrarian society. Again we see the plea for the complexity of “*wicked problems*” to be solved.

The downside is that those planning processes of the second-generation are time consuming and need a careful and tedious procedure in developing the most important and difficult *zero-step* – before such decision making can be started. It also implies an exchange of knowledge between the parties beforehand, in order to minimize *hidden agendas*. It also must be emphasized that those decision-making processes do not lead necessarily to a predefined goal, they are often *open-ended* and demand flexibility among the discourse participants, who need to remain open-minded.

The more questions we ask the more answers are possible and vice versa. Limitations of technological solutions are always hidden in the open ecological and social systems: Just compare the (in)famous case of DDT sprayings in the past [413–415]. Today, it is clear that with linear planning, DDT has been banned for ecological and health reasons, not considering the wider argument field of malaria prophylaxes. This inconsiderate DDT ban has caused millions of malaria deaths in Africa. Today, reasonable domestic use of DDT has again lowered the malaria threat measurably.

Constraints in possible secondary effects in ecology should be examined carefully. This is well demonstrated in the case of the Monarch larvae being killed by Bt-Maize-Pollen, the result of a laboratory study published in *Nature* [416] where the subsequent press interpretation got way out of proportion – even though the author Losey himself warned about the limitations of this small lab study. Would researchers have asked the farmers, they would have been able to say that feeding time of the young larvae do rarely overlap with the time of pollen shed of maize, and that the plants the Monarchs are feeding upon are fiercely fought as a weed. Subsequent field studies revealed that there is no problem arising from extensive Bt maize planting for the Monarch larvae [12].

In order to tackle wicked problems, you need to go through *an extensive process of argumentation*, also called objectification, not to be mixed up with an “objective approach” to the problem.

There is rational planning, but there is no way to start to be rational: One should always start a step earlier, since there are important trends and facts which will make straightforward rational thinking and acting in solving wicked problems useless. It is not the theory component, but rather the political component of the knowledge, which determines the vector of the action. This is the *zero-step* so important in the publications of Horst Rittel [121, 122].

As an example: The fact, that experts can be wrong and farmers know better in certain situations in agriculture because they are better observers out in the field and because they are very experienced in traditional knowledge [417].

The knowledge needed in solving wicked planning problems is not concentrated in a single head. It is absolutely essential to let all partners be involved in the problem solution process, which includes part of the population (mainly farmers’ organizations and consumer organizations), the Governmental Regulators, the Non-Governmental Organisations, the Life Science Companies, and the Scientists. There is no monopoly of knowledge. Having illustrated the difficulties in solving wicked problems, we need a new approach in problem solving, in order to avoid the pitfalls of ignoring bottom up feedbacks.

You only can keep to this rule if you are also following another important rule. All partners in the

planning process have to avoid hidden agendas, which is certainly eased by a minimum amount of respect paid to each other partner. Nobody should be criticized for speaking up in his own interest.

A caveat: It would be naive to just believe in the discursive capacities of the civil society, contrary to what Gerhards [418] has shown – that Habermas’ support for the discursive model is based on the assumption that actors of the civil society argue much more discursively and on a higher level of rationality than other collective actors do. But empirical results show that actors of the civil society are, maybe, even less discursive than other actors.

It is primarily the paradox of rationality which has been severely underestimated in the systems approach of the first generation when tackling *wicked problems*.

How to Solve Wicked Problems in Biotechnology and the Environment

What we need in such cases is an action-oriented approach. Risk Assessment and Management must be seen as a planning strategy of the second generation in developing a professional framework for *decision making*.

Strategies have to be developed to recognize the consequences of our doing on one side, and to specify our knowledge on the other side. This knowledge has to be gained step by step and case by case. If we want to clearly distinguish our present state knowledge from appropriate decisions to be made *not* based on our views and opinions, we need to go through the following steps:

- What is the problem?
- What do we want?
- What are the alternatives?
- How do we compare them?
- How can we reach the solution?

All participants need to keep in mind that there are *various types of planning knowledge* (arranged according to the five questions asked above).

Examples given here are lumped together as simple keyword illustrations, taken out of their context in real planning examples, and they cannot be regarded as an example of a realistic situation; this would be exactly the task of a planning process of the second generation.

Factual knowledge is the knowledge of what actually happens (quantitative data or empirical, observational

data). Gene flow species by species/region by region/facts about insect resistance in agriculture.

- *Deontic Knowledge*, the very important knowledge of what ought to be. The knowledge about new crops which enhance agricultural production/new agricultural techniques to avoid erosion/new biological approaches to fight insect pests etc.
- *Explanatory Knowledge* explains why things are so or why certain effects will happen. Here, you already start to determine the direction of the solution. The way Bt proteins are acting on specific pest and beneficial insects/what are the main reasons of unwelcome erosion effects/mechanisms of vertical gene flow/mechanisms of resistance development.
- *Instrumental knowledge* on how to steer certain processes, on how to achieve certain goals, knowledge which needs to be balanced against regulation and safety. The way how to build Bt and other genes into crops and how to stabilize them/how to avoid vertical gene flow/how to avoid unwelcome soil erosion/how to avoid early upcoming pest resistance.
- *Conceptual knowledge* which would allow avoiding conflicts before they pop up. This is the knowledge about complex situations, taking into account all previous kinds of knowledge and also weighing them against arguments coming from open ecological and societal systems. Concepts about transgenic crops compatible to the ideas of a sustainable agriculture. Lawyers and judges also may work with this kind of procedural knowledge.

You need to go through an *extensive, time-consuming process of argumentation*, also called objectification, not to be mixed up with an “objective approach” to the problem. The hopes of this process are:

- To forget less, to raise the right issue
- To look at the planning process as a sequence of events
- To stimulate doubt by raising questions, to avoid short-sighted explicitness
- To control the delegation of judgment. Experts have no absolute power; scientific knowledge is important, but always limited.

There is no such thing as “scientific planning.”

- Solving practical problems as to develop sustainable transgenic crops cannot be dealt with by

“scientification of planning.” Dealing with wicked problems is always political because of its deontic premises (means that you have to involve knowledge what ought to be) and because we deal with traditional knowledge. Science only generates factual, instrumental, and in the best case explanatory knowledge.

- The planner (here the manager of an action plan) is not primarily an expert, but a *mid-wife of problem solving*, a teacher more than a doctor. Moderate optimism and careful seasoned disrespect, casting doubt is a virtue, not a disadvantage of an action plan manager.
- The planning process of wicked problems has to be understood as an *argumentative process*, it should be seen as a venture (or even *adventure*) within a conspiracy framework, where one cannot anticipate all the consequences of plans.
- Systems methods of the *second generation* are trying to make this deliberation explicit, to support it and to find means in order to make this process more powerful and to get it under better control *for all participants*. Methods like the computer-based argument mapping systems of can be helpful [419].
- It helps making such processes more successful if they are conducted in the spirit of the *Symmetry of Ignorance* [420] – this is the secret of the active listening which often leads to acceptable outcomes and trust.

This seems to be a rather theoretical approach with lots of restrictive rules, but actually it is, on the contrary, an opening for much more freedom in dialogue. Also, it is more practical and efficient in creating results and contrasts with the traditional stakeholder concept where hidden agendas prevail in often disguised authoritarian structures. Such discursive processes are described in detail [80, 121–123, 421–425]. A comprehensive and voluminous monograph on risk-related debate methods has been published by Ortwin Renn [426], see especially the texts related to risk communication with essays 7 and 8 and section 8 on risk participation with numerous references, but notably lacking completely the papers on the “Systems Approach” of the Churchman/Rittel/Webber school.

In a French paper, the origin of negatively connoted words in the debate on GM crops like “contamination,”

“pollution,” “Frankenfood,” etc., Moirand [427] clearly reveals the links to negative events like BSE, dioxin scandals, and of course Tchernobyl, etc., thus explaining new words like “mad soya” and “mad colza” in the media. Moirand concludes that a new type of discourse is needed, but also Renn [426] does not refer to the very pragmatic and promising systems approach of Churchman and Rittel.

There are many more schools promoting discourse and new decision-making processes, also in specialized journals, only a few can be summarized here for space reason: [75, 76, 78, 84, 119, 120, 411, 427–441].

See Patrick Moore’s practical examples of decision-making processes solving environmental and sustainability problems in forestry, consult his own website Green Spirit <http://www.greenspirit.com/index.cfm>. These processes need time. Patrick Moore [442–444] has gone successfully through such processes in the difficult task of reconciliation between the needs of timber production and environmental constraint; he needed months of debate to come to reasonable decisions.

Another good example on how group discourses have good learning effects, has been described by Snyder et al. [258]: Although the US government has assured stakeholders of their safety, the EU continues to be an outspoken opponent. This can largely be attributed to a lack of trust in the regulatory process, and especially a cynical perspective on the underlying science and institutions that govern approval. Such disparities were illustrated in 2003 when the USA donated GM maize to aid African countries stricken by famine. Under purported EU threats, negative propaganda by NGOs, and stressing retaliatory trade sanctions, African officials refused the aid. An examination of this episode contrasts the potential discord between those affected and those who formulate government policy. Using resources from both sides of the debate, this scenario summarizes the pertinent issues regarding EU’s refusal to the import of transgenic crops. A group discussion and debate protocol was developed for facilitating small group and entire class consideration of the scenario while strengthening student critical thinking skills.

It helps, if you prepare carefully scenarios before people start the process, a method which has been successfully applied to the reconciliation processes in

South Africa after abolishing apartheid by Adam Kahane, one of the principal mediators [445]. He also followed another wise rule: Should only people participate in such processes who are part of the problem. Another excellent example of long-term discourse is described in many aspects by von Grebmer et al. [437]:

- By working collectively the process will be more open, transparent, inclusive and accountable, and sensitive to the normative dimensions of the issues critical to the participants. The themes and processes outlined in this article set the stage for the discussions, internally and between countries, that will shape the policies of agricultural biotechnology in the region. If the dialogue can frame the discussion and be enriched by the information generated from actions taken, it can sustain the interest and commitment of the stakeholders, and more successfully direct biotechnology toward reducing hunger and poverty in the region.

There are too many scientists remaining in the ivory tower, shying away from public debates. They fear losing their independence, a fear which is not just unfounded, but actually it is the contrary: remaining in the academic ivory tower means having lost your independence, since science is not an art per se, it is full of importance for society and humanity. A strong plea in this direction is coming from [446]. Although science should remain at the heart of invention and the drive to make our lives better, scientists should, instead of always having “the answers” ready, should not be afraid to engage in a contradictory evidence-based mode.

In one of the most successful examples of long-term discourse, the author participated as an invited expert in a public hearing in 2000. Strikingly, it was done without the theoretical load described above, but with lots of financial and logistic help from the New Zealand Government, in particular from the Royal Commission on Genetic Modification. A report was finalized after a 14-month inquiry into the risks and benefits of genetic modification. It heard from over 400 experts, including scientists, environmentalists, and ethical specialists. It considered more than 10,000 public submissions and heard the view of many others during a series of public meetings, hui, and workshops around New Zealand.

The Royal Commission’s major conclusion was that New Zealand should proceed cautiously with genetic

modification (GM) but not close the door to the opportunities offered by the new technology <http://www.mfe.govt.nz/issues/organisms/index.html>. The discourse is still continuing. Again, it is visible that the discourse is less confrontational and may lead to innovative solutions in the future [447]:

- The debate about genetic modification (GM) can be seen as characteristic of our time. Environmental groups, in challenging GM, are also challenging modernist faith in progress, and science and technology. In this paper we use the case of New Zealand's Royal Commission on Genetic Modification to explore the application of science discourses as used by environmental groups. We do this by situating the debate in the framework of modernity, discussing the use of science by environmental groups, and deconstructing the science discourses evident within environmental groups' submissions to the Commission. We find science being called into question by the very movement that has relied on it to fight environmental issues for many years. The environmental groups are challenging the traditional boundaries of science, for although they use science they also present it as a culturally embedded activity with no greater epistemological authority than other knowledge systems. Their discourses, like that of the other main actors in the GM debate, are thus part of the constant re-negotiation of the cultural construct of 'science'.

However, this process should not be mollified on the costs of hard science. The line between science and pseudoscience is often difficult to draw.

A Remark About the Psychology of the GMO Debate To be written in the next coming days.

It should also be possible to think and act in relation to the reconciliation of science and spirituality, since it will be an important element besides the ratio of science, the ethics of our societal activities, and the emotional elements in human life. But it will be difficult to separate the cheap esoteric chaff from the precious seeds of true spirituality, as Helmut Reich's writings demonstrate [448]. We must endeavor new fields of thought, as done by Papazova Ammann [449], a Bulgarian-born Swiss philosopher with roots in the schools of Muntjan and Rittel.

- What do we need as visionaries: Progress or Development? This is my question today, as I deal with the topic of Biovisionaries here in the Library of Alexandria. I ask this question because I am convinced that we need to build a new culture of questioning. We need a culture orienting itself by authentic questions. How can we develop taste and the ability to distinguish between those questions which are cognitive, statement-oriented and those which are authentic, close to life and to people? What is more important: cognizance or decision for action? How can we move between Statements and Questions? Statements reflect the need to understand the world. But they are the result of past experience and are often contained in frameworks which are coined by society. They may even protect old routines which hinder innovation. Questions, in contrast to statements, can transform our judgements and prejudices. Questions give birth to energy for new orientation, for a more conscious future. This orientation towards the future, towards vision provokes those choice-questions, and they alone will open the way for an urge to change the world. Visions need people who are free! The quality of freedom is inherent in the question. We must strive for this quality through choice-questions. If we cannot befriend these choice-questions with science, it will disengage from the questioners and will not be human science anymore. Thus we need a new humility of thinking – as it has been wonderfully defined by the German philosopher Heidegger: "The question is the devoutness of thinking".

Conclusions Only a multifaceted dialogue over a considerable time span will lead to success. The Internet scene is developing fast and new communication software tools are available now, so careful scrutiny for such a network of networks need to be done first, and the big players like Google and competing networks should be consulted as well.

Personal experience in dialogue with many networkers reveals that sometimes important networks are only known in specific clusters, these lacunas should be closed for many reasons – see section [Illusions and Realities on Educational Effects in the Debate, the Dialogue Between Science and the Public](#). Knowledge exchange, jumping over national fences, and coordination will be a follow-up effect, without

even declaring it to be the goal of such activity. As for now, this is just an idea and needs to be discussed with Internet and website specialists. After all, the leading webmasters and coordinators agree that it is time to *enhance collaboration through better communication*.

ASK-FORCE can contribute to this process in making sure that professional peer-reviewed risk assessment papers are fed into the dialogue processes and are ideally fed into a life decision-making process with relevant participants.

Bibliography

1. Beardmore JA (1997) Transgenics: autotransgenics and allotransgenics. *Trans Res* 6(1):107–108
2. Taverniers I et al (2008) Gene stacking in transgenic plants: towards compliance between definitions, terminology, and detection within the EU regulatory framework. *Environ Biosaf Res* 7(4). doi:10.1051/eb:2008018
3. Potrykus I et al (2010) Transgenic plants for food security in the context of development, statement of the pontifical academy of sciences. *New Biotechnol* 27(5):443–717
4. Rauschen S (2009) German GM research – a personal account. *Nat Biotech* 27(4):318–319
5. Showalter E (1997) *Hystories*. Columbia University Press, New York, 244 pp
6. Linden A, Fenn J (2003) Understanding Gartner's hype cycles. In: *Strategic Analysis Report*. Gartner Research, p 12
7. Chassy BM (2007) The history and future of GMOs in food and agriculture. *Cereal Foods World* 52(4):169–172
8. Martineau B (2002) *First fruit: the creation of the Flavr savr tomato and the birth of biotech food*. McGraw-Hill, New York, 224 pp
9. Carrière Y, Crowder DW, Tabashnik BE (2010) Evolutionary ecology of insect adaptation to Bt crops. Blackwell, Oxford, pp 561–573
10. Ellstrand NC et al (2010) Crops gone wild: evolution of weeds and invasives from domesticated ancestors. Blackwell, Oxford, pp 494–504
11. Huang F, Andow DA, Buschman LL (2011) Success of the high-dose/refuge resistance management strategy after 15 years of Bt crop use in North America. *Entomol Experiment Et Appl* 140(1):1–16
12. Gatehouse AMR, Ferry N, Raemaekers RJM (2002) The case of the monarch butterfly: a verdict is returned. *Trends Genet* 18(5):249–251
13. Lovei GL, Andow DA, Arpaia S (2009) Transgenic insecticidal crops and natural enemies: a detailed review of laboratory studies. *Environ Entomol* 38:293–306
14. Shelton A et al (2009) Setting the record straight: a rebuttal to an erroneous analysis on transgenic insecticidal crops and natural enemies. *Trans Res* 18(3):317–322
15. Moore GA (2002) *Crossing the chasm*. Harper Paperbacks, New York, Revised edition 20 Aug 2002, 256 pp
16. Thro AM (2004) Europe on transgenic crops: how public plant breeding and eco-transgenics can help in the transatlantic debate. *Commentary. AgBioForum* 7:142–148
17. Adams J (1995) *Risk*. Taylor & Francis, Bristol, 228 pp
18. Royal-Society (2009) Reaping the benefits: science and the sustainable intensification of global agriculture. In: *RS Policy document 11/09*. Royal Society, London, p 89
19. Arber W (2010) Genetic engineering compared to natural genetic variations. *New Biotechnol* 27(5):517–521
20. Britt AB, May GD (2003) Re-engineering plant gene targeting. *Trends Plant Sci* 8(2):90–95
21. Henderson IR, Jacobsen SE (2007) Epigenetic inheritance in plants. *Nature* 447(7143):418–424
22. Johnson L (2007) The genome strikes back: the evolutionary importance of defence against mobile elements. *Evolut Biol* 34(3):121–129
23. Moch K, Brauner R, Ott B (2005) Epigenetics, transgenic plants & risk assessment. In: *Epigenetics, transgenic plants & risk assessment*. 1st Dec 2005, Literaturhaus, Frankfurt am Main, Germany © 2006, Öko-Institut e.V., Box 50 02 40, D-791028 Freiburg, Die Deutsche Bibliothek – CIP Cataloguing-in-Publication-Data, A catalogue record for this publication is available from Die Deutsche Bibliothek
24. Smilde AK et al (2010) Dynamic metabolomic data analysis: a tutorial review. *Metabolomics* 6(1):3–17
25. Domon B, Aebersold R (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol* 28(7):710–721
26. Addona TA et al (2009) Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol* 27(7):633–U85
27. Wittenberg AHJ et al (2005) Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. *Molec Genet Genomics* 274(1):30–39
28. Colbert T et al (2001) High-throughput screening for induced point mutations. *Plant Physiol* 126(2):480–484
29. Giddings VL (2006) "Cisgenic" as a product designation. *Nat Biotech* 24(11):1329–1329
30. Schouten HJ, Jacobsen E (2007) Are mutations in genetically modified plants dangerous?. *J Biomed Biotechnol*: 8261, p 2
31. Schouten HJ, Krens FA, Jacobsen E (2006) Do cisgenic plants warrant less stringent oversight? *Nat Biotechnol* 24(7):753–753
32. Schouten HJ et al (2006) Cisgenic plants are similar to traditionally bred plants – International regulations for genetically modified organisms should be altered to exempt cisgenesis. *Embo Reports* 7(8):750–753
33. Jacobsen E, Nataraja KN (2008) Cisgenics – Facilitating the second green revolution in India by improved traditional plant breeding. *Curr Sci* 94(11):1365–1366
34. Jacobsen E, Schouten HJ (2007) Cisgenesis strongly improves introgression breeding and induced translocation breeding of plants. *Trends Biotechnol* 25(5):219–223
35. Conner AJ et al (2007) Intragenic vectors for gene transfer without foreign DNA. *Euphytica* 154(3):341–353

36. Rigola D et al (2009) High-throughput detection of induced mutations and natural variation using keypoint (TM) technology. *PLoS One* 4(3):e4761
37. Parry MAJ et al (2009) Mutation discovery for crop improvement. *J Exper Botany* 60(10):2817–2825
38. Davies H, Bryan G, Taylor M (2008) Advances in functional genomics and genetic modification of potato. *Potato Res* 51(3):283–299
39. Townsend JA et al (2009) High-frequency modification of plant genes using engineered zinc-finger nucleases. *Nature* 459(7245):442–445, advanced online publication
40. Shukla VK et al (2009) Precise genome modification in the crop species *Zea mays* using zinc-finger nucleases. *Nature* 459(7245):437–441, advanced online publication
41. Cai C et al (2009) Targeted transgene integration in plant cells using designed zinc finger nucleases. *Plant Molec Biol* 69(6):699–709
42. Osakabe K, Osakabe Y, Toki S (2010) Site-directed mutagenesis in *Arabidopsis* using custom-designed zinc finger nucleases. *Proc Nat Acad Sci* 107(26):12034–12039
43. Gabriel R et al (2011) An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat Biotech* 29(9): 816–823
44. Ammann K (2008) Feature: integrated farming: why organic farmers should use transgenic crops. *New Biotechnol* 25(2):101–107
45. Mahfouz MM et al (2011) De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proc Nat Acad Sci* 108(6):2623–2628
46. CNBS (2011) Plant genomics “Molecular scissors” developed at KAUST. CNBS, PR newswire. DOI: http://www.cnbc.com/id/41731207/Plant_Genomics_Molecular_Scissors_Developed_at_KAUST and <http://www.ask-force.org/web/Genomics/CNBC-Kaust-Genomic-Scissors-2011.PDF>
47. Epinat JC et al (2003) A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucl Acid Res* 31(11):2952–2962
48. Paques F, Duchateau P (2007) Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. *Curr Gene Ther* 7(1):49–66
49. Silva G et al (2011) Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Curr Gene Ther* 11(1):11–27
50. Benner SA (2004) Understanding nucleic acids using synthetic chemistry. *Acc Chem Res* 37(10):784–797
51. Tian JD, Ma KS, Saaem I (2009) Advancing high-throughput gene synthesis technology. *Molec Biosyst* 5(7):714–722
52. Benner SA et al (1998) Redesigning nucleic acids. *Pure Appl Chem* 70(2):263–266
53. Benner SA, Sismour AM (2005) Synthetic biology. *Nat Rev Genet* 6(7):533–543
54. Rusch DB et al (2007) The sorcerer ii global ocean sampling expedition: Northwest Atlantic through Eastern tropical pacific. *Plos Biol* 5(3):398–431
55. Gibson DG et al (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319(5867):1215–1220
56. Gibson DG et al (2008) One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome. *Proc Nat Acad Sci USA* 105(51):20404–20409
57. Gibson DG et al (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329(5987):52–56. doi:10.1126/science.1190719
58. Bugl H et al (2007) DNA synthesis and biological security. *Nat Biotechnol* 25:627–629
59. Maurer SM, Lucas KV, Terrell S (2006) From understanding to action: community-based options for improving safety and security in synthetic biology, Draft 1.1, 15 April 2006. University of California, Berkeley, California, p 93
60. Serrano L (2007) Synthetic biology: promises and challenges. *Mol Syst Biol* 3:158
61. Edmond G, Mercer D (2009) Norms and irony in the biosciences: ameliorating critique in synthetic biology. *Law Literat* 21(3):445–470
62. Miller HI (2010) Understanding the frankenstein myth. Project Syndicate, a World of Ideas. <http://www.ask-force.org/web/Genomics/Miller-Understanding-Frankenstein-Tradition-2010.pdf>
63. Tabashnik BE et al (2011) Efficacy of genetically modified Bt toxins against insects with different genetic mechanisms of resistance. *Nat Biotech*, advance online publication
64. Soberon M et al (2007) Engineering modified Bt toxins to counter insect resistance. *Science* 318(5856):1640–1642
65. Keller D (2009) Start talking to each other! – dialogue as key to biotechnology’s future in Europe. *New Biotechnol* 25:185, Corrected proof
66. Ramon D, Diamante A, Calvo MD (2008) Food biotechnology and education. *Electron J Biotechnol* 11(5)
67. Harms U (2002) Biotechnology education in schools. *Electron J Biotechnol* 5(3):205–211
68. Gensuisse (2011) Themenfokus. Gensuisse Website Forum 2011. Available from: <http://www.gensuisse.ch/focus/index.html>
69. Sengooba T et al (2009) Biosafety education relevant to genetically engineered crops for academic and non-academic stakeholders in East Africa. *Electron J Biotechnol* 12(1)
70. McHughen A (2007) Public perceptions of biotechnology. *Biotechnol J* 2(9):1105–1111
71. James C (2009) Global status of commercialized biotech/GM crops. Brief 39, Executive Summary. ISAAA, p 20
72. Sturgis P, Allum N (2004) Science in society: re-evaluating the deficit model of public attitudes. *Public Understand Sci* 13(1):55–74
73. Sturgis P, Cooper H, Fife-Schaw C (2005) Attitudes to biotechnology: estimating the opinions of a better-informed public. *New Genet Soc* 24(1):31–56
74. Sturgis P, Roberts C, Allum N (2005) A different take on the deliberative poll – Information, deliberation, and attitude constraint. *Public Opin Quart* 69(1):30–65

75. Gaskell G et al (2000) Biotechnology and the European public. *Nat Biotechnol* 18(9):935–938
76. Schuman H, Presser S (1980) Public-opinion and public ignorance – the fine line between attitudes and non-attitudes. *Am J Sociol* 85(5):1214–1225
77. Aerni P, Scholderer J, Ermen D (2011) How would Swiss consumers decide if they had freedom of choice? Evidence from a field study with organic, conventional and GM corn bread. *Food Policy*
78. Irwin A (2006) The politics of talk: coming to terms with the “new” scientific governance. *Soc Stud Sci* 36(2):299–320
79. Churchman CW (1979) The systems approach and its enemies and (Commented German transl.: *Der Systemansatz und seine “Feinde,”* with an Introduction by the editor and translator, Werner Ulrich, ed. and transl., Paul Haupt, Bern, 1981). Basic Books, New York
80. Rittel HWJ, Webber MR (2005) Dilemmas in a general theory of planning. *Policy Sci* 4(2):155–169
81. Rittel H (1992) Planen, entwerfen, design, ausgewählte schriften zu theorie und methodik. In: Reuter Wolf D (ed) *Planen, entwerfen, design*. Verlag W. Kohlhammer, Berlin, p 432
82. Protzen JP, Harris DW (2010) *The universe of design*: Horst Rittel’s theories of design and planning, 1st edn. Routledge, London/New York, p 264, 19 June 2010
83. Magnan A (2003) Refeudalizing the public sphere: “Manipulated publicity” in the Canadian debate on GM foods. In: Annual meeting of the canadian-sociology-and-anthropology-association (CSAA). University of Alberta, Halifax, Canada
84. Vaughan E (1995) The significance of socioeconomic and ethnic diversity for the risk communication process. *Risk Anal* 15(2):169–180
85. Osseweijer P (2006) A new model for science communication that takes ethical considerations into account – The three-E model: entertainment, emotion and education. *Sci Eng Ethics* 12(4):591–593
86. Osseweijer P (2006) Imagine projects with a strong emotional appeal. *Nature* 444(7118):422–422
87. Osseweijer P (2006) A short history of talking biotech, fifteen years of iterative action research in institutionalising scientists’ engagement in public communication. *Vrije Universiteit, Amsterdam*
88. Osseweijer P, Ammann K, Kinderlerer J (2010) Societal issues in industrial biotechnology. In: Soethaert W, Vandamme EJ (eds) *Industrial biotechnology, sustainable growth and economic success, handbook*. Wiley, VCH Verlag, Weinheim, pp 457–481, Chapter 14, 522 pp
89. Koutsogiannis D, Mitsikopoulou B (2004) The Internet as a glocal discourse environment – A commentary on “second language socialization in a bilingual chat room” by Wan Shun Eva Lam and “second language cyber rhetoric: A study of Chinese L2 writers in an online usenet group” by Joel Bloch. *Lang Learn Technol* 8(3):83–89
90. Kostoff RN et al (2006) The structure and infrastructure of the global nanotechnology literature. *J Nanoparticle Res* 8(3–4):301–321
91. Kanter RM (2000) Are you ready to lead the e-cultural revolution? *Inc* 22(2):43–44
92. Bruns A (2008) *Blogs. Wikipedia. Second life and beyond (digital formations)*. Peter Lang, Bern
93. Reifer D (2002) Ten deadly risks in internet and intranet software development. *IEEE Software* 19(2):12–14
94. Kalman ME et al (2002) Motivations to resolve communication dilemmas in database-mediated collaboration. *Commun Res* 29(2):125–154
95. Borland N, Wallace D (1999) Environmentally conscious product design: a collaborative internet-based modeling approach. *J Indust Ecol* 3(2–3):33–46
96. Dall’Olio GM et al (2011) Ten simple rules for getting help from online scientific communities. *PLoS Comput Biol* 7(9): e1002202
97. Degraffi G, Alexandrova N, Ripandelli D (2003) Databases on biotechnology and biosafety of GMOs. *Environ Biosaf Res* 2(3):145–160
98. Burns CG (2011) Biosafety resources on the Internet. *J Int Wildlife Law & Policy* <http://www.jiwl.com/> 20110801. Available from: http://www.jiwl.com/contents/biosafety_resources_net.html
99. Ammann K (2011) List of websites related to GM crops and biotechnology. DOI: <http://www.ask-force.org/web/Sustainability/Websites-List-Publ.def.pdf>
100. Leydesdorff L (2002) Indicators of structural change in the dynamics of science: entropy statistics of the SCI journal citation reports. *Scientometrics* 53(1):131–159
101. Leydesdorff L (2008) Caveats for the use of citation indicators in research and journal evaluations. *J Am Soc Inform Sci Technol* 59(2):278–287
102. Leydesdorff L (2009) How are new citation-based journal indicators adding to the bibliometric toolbox? *J Am Soc Inform Sci Technol* 60(7):1327–1336
103. Leydesdorff L, Wagner C (2009) Macro-level indicators of the relations between research funding and research output. *J Informet* 3(4):353–362
104. Leydesdorff L, Hellsten I (2006) Measuring the meaning of words in contexts: An automated analysis of controversies about “Monarch butterflies”, “Frankenfoods”, and “stem cells”. *Scientometrics* 67(2):231–258
105. Aizen J et al (2004) Traffic-based feedback on the web. *Proc Nat Acad Sci USA* 101(Suppl 1):5254–5260
106. Cavaller V (2009) Scientometrics and patent bibliometrics in RUL analysis: a new approach to valuation of intangible assets. *Vine* 39:80–91
107. Cavaller V, Aubertin C (2008) Elements of scientometrics and patent bibliometric-analysis for the estimated remaining useful life (RUL) in the valuation of intangible assets. In: *Proceedings of the 5th international conference on intellectual capital and knowledge management and organisational learning*, New York, pp 87–95
108. Laporte RE et al (2002) Papyrus to powerpoint (P 2 P): metamorphosis of scientific communication. *Brit Med J* 325(7378):1478–1481

109. Sa ER et al (2003) Open source model for global collaboration in higher education. *Int J Med Inform* 71(2–3):165–165
110. Linkov F et al (2003) Globalisation of prevention education: a golden lecture. *Lancet* 362(9395):1586–1587
111. Linkov F, The I (2006) Internet-based supercourse system. *J Public Health Policy* 27(4):442–443
112. Laporte RE et al (2002) Infopoints – Whisking research into the classroom. *Brit Med J* 324(7329):99–99
113. Laporte RE et al (2006) A scientific supercourse. *Science* 312(5773):526–526
114. Sauer F, Bennett S, Cha M, Linkov F, LaPorte R (2010) Supercourse, Bibliotheca Alexandrina, and the educator as catalyst. *Educause Quart* 33(3)
115. Ammann K (2011) Presentations for conferences etc. with powerpoint slides. In Audio-Visual Material 20110904, Ammann K, Neuchatel
116. Adly N (2009) Bibliotheca alexandrina: a digital revival. *Educause Rev* 44(6):8–9
117. Craig W et al (2008) An overview of general features of risk assessments of genetically modified crops. *Euphytica* 164(3):853–880
118. Leicht EA, Newman MEJ (2008) Community structure in directed networks. *Phys Rev Lett* 100(11):118703
119. Newman MEJ (2003) The structure and function of complex networks. *Siam Rev* 45:167–256
120. Saner M (2007) A map of the interface between science & policy, staff papers. Council of Canadian Academies, Ottawa, p 15
121. Rith C, Dubberly H (2007) Horst W. J. Rittel's writings on design: select annotations. *Des Issues* 23(1):75–77
122. Rittel H, Weber M (1973) Dilemmas in a general theory of planning. *Policy Sci* 4:155–169
123. Ammann K, Papazova Ammann B (2004) Factors influencing public policy development in agricultural biotechnology. In: Shantaram S (ed) Risk assessment of transgenic crops. Wiley, Hoboken, p 1552
124. Gasson M, Burke D (2001) Scientific perspectives on regulating the safety of genetically modified foods. *Nat Rev Genet* 2(3):217–222
125. Phillips PWB (2003) Traceability and trade of genetically modified food. *Biotechnol Sci Soc Crossroad* 5:141–154
126. Sheehy RE, Kramer M, Hiatt WR (1988) Reduction of polygalacturonase activity in tomato fruit by antisense rna. *Proc Nat Acad Sci USA* 85(23):8805–8809
127. Redenbaugh K et al (1994) Regulatory Assessment of the Flavr-savr tomato. *Trends Food Sci Technol* 5(4):105–110
128. Kramer MG, Redenbaugh K (1994) Commercialization of a tomato with an antisense polygalacturonase gene – the Flavr Savr(Tm) tomato story. *Euphytica* 79(3):293–297
129. Krieger EK et al (2008) The Flavr Savr tomato, an early example of RNAi technology. *Hortscience* 43(3):962–964
130. Graff G, Zilberman D (2004) Explaining Europe's resistance to agricultural biotechnology. *Agric Resour Econ* 7(5):4
131. Lawrence F (2009) It is too late to shut the door on GM foods consumers said no to the GM farming giants a decade ago, but that didn't stop millions of tonnes of their soya entering the food chain, in The Guardian. The Guardian and Observer, London
132. Flachowsky G et al (2007) Studies on feeds from genetically modified plants (GMP) – Contributions to nutritional and safety assessment. *Anim Feed Sci Technol* 133(1–2):2–30
133. Aumaitre A (2004) Safety assessment and feeding value for pigs, poultry and ruminant animals of pest protected (Bt) plants and herbicide tolerant (glyphosate, glufosinate) plants: interpretation of experimental results observed worldwide on GM plants. *Italian J Anim Sci* 3(2):107–121
134. Paarlberg R (2006) Are genetically modified (GM) crops a commercial risk for Africa? *Int J Technol Globalisation* 2(1–2):81–92
135. Cohen JI, Paarlberg R (2002) Explaining restricted approval and availability of GM crops in developing countries. *AgBiotechNet* 4:1–6
136. Gruere GP, Carter CA, Farzin YH (2008) What labelling policy for consumer choice? The case of genetically modified food in Canada and Europe. *Canad J Econom-Revue Canadienne D Economique* 41(4):1472–1497
137. Gruere GP, Rosegrant MW (2008) Assessing the implementation effects of the biosafety protocol's proposed stringent information requirements for genetically modified commodities in countries of the Asia Pacific economic cooperation. *Rev Agric Econom* 30(2):214–232
138. Gruere GP, Sengupta S (2009) Biosafety decisions and perceived commercial risks, The role of GM-free private standards. In: IFPRI Discussion Paper 00847, Environment and Production Technology Division. FPRI, Washington DC, p 40
139. Greenpeace (2007) Contamination Report 2006, annual review of cases of contamination, illegal planting and negative side effects of genetically modified organisms. Greenpeace International, Amsterdam, p 24
140. Greenpeace (2008) Contamination Report 2007, annual review of cases of contamination, illegal planting and negative side effects of genetically modified organisms. Greenpeace International, Amsterdam, p 48
141. ISAAA (2011) Cotton (*Gossypium hirsutum* L.) events. ISAAA 2011 11. Oct 2011. Available from: <http://www.isaaa.org/gmaprovaldatabase/cropevents/default.asp?CropID=6>
142. Sadashivappa P, Qaim M (2009) Bt cotton in India: development of benefits and the role of government seed price interventions. *AgBioForum* 12:172–183
143. Mueller-Jung J (2007) Wie verpackt man eine Kulturrevolution in Watte? How to wrap up a cultural revolution in cotton wool?. In: Frankfurter Allgemeine Zeitung. Frankfurt. p N1
144. Gruere G, Meththa-Bhatt P, Sengupta D (2008) Bt cotton and farmer suicides in India, reviewing the evidence. IFPRI-Discussion Paper 2008, 00808
145. Gruere G, Sengupta D (2011) Bt cotton and farmer suicides in India: an evidence-based assessment. *J Develop Stud* 47(2):316–337
146. Shiva V (2004) The suicide economy of corporate globalisation. *Z Net – The spirit of resistance lives 2004*. Available from:

- <http://www.zcommunications.org/the-suicide-economy-of-corporate-globalisation-byvandana2-shiva>
147. Sunilkumar G et al (2006) From the cover: engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. *Proc Natl Acad Sci* 103(48):18054–18059. doi:10.1073/pnas.0605389103
 148. Choudhary B, Gaur K (2011) Bt cotton in India, a multipurpose crop. In: ISAA (ed) Celebrating 10 years. International Service for the Acquisition of Agri-Biotech Applications, Biotech Information Center, New Delhi, p 6
 149. Graff G, Hochman G, Zilberman D (2009) The political economy of agricultural biotechnology policies. *AgBioForum* 12:1–13, <http://www.agbioforum.org/v12n1/v12n1a04-graff.htm> and <http://www.botanischergarten.ch/Regulation/Graff-Political-Economy-Policies-2009.pdf>
 150. Ayal S, Hochman G (2009) Ignorance or integration: the cognitive processes underlying choice behavior. *J Behav Decis Mak* 22(4):455–474
 151. Ministerio da Ciencia e Tecnologia (2005) CTN Bio, Biosafety Law Nº 11.105, of 24 March 2005. Ministerio da Ciencia e Tecnologia, Brazilia, p 17
 152. European Parliament and European Council (2003) Regulation (EC) No 1829/2003. *Off J Euro Union L* 268(1):1–23, 20030922
 153. The European parliament and the council of the European union (2010) EU-Regulation-GMO-free regions, GMOs: Member states to be given full responsibility on cultivation in their territories, IP/10/921. 20100713, The European parliament and the council of the European union, Brussels, p 2
 154. James C (2009) Global status of commercialized biotech/GM Crops. In: ISAA (ed) AAA briefs. The International Service for the Acquisition of Agri-biotech Applications (ISAAA), Ithaca
 155. Galvao A (2010) *Celeres*, Biotechnology Report 2010. 20100809, Uberlandia, Matto Grosso, Celeres, p 7
 156. Marques R, Neto CGA (2007) The Brazilian system of innovation in biotechnology: a preliminary study. *J Technol Manag Innov* 2(1):55–63
 157. Mendonca-Hagler L et al (2008) Trends in biotechnology and biosafety in Brazil. *Environ Biosaf Res* 7(3):115–121
 158. Silveira JM, Ferreira J, Dal Poz ME, Alssad A (2004) *Biotecnologia e recursos genéticos: desafios e oportunidades para o Brasil*/Biotechnology and genetic resources: challenges and opportunities for Brazil. Campinas Instituto de Economia, Rio de Janeiro, p 412
 159. Frohme M et al (2000) Mapping analysis of the *Xylella fastidiosa* genome. *Nucl Acids Res* 28(16):3100–3104
 160. Simpson AJG et al (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* 406(6792):151–157
 161. Vasconcelos A (2003) The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. *Proc Nat Acad Sci USA* 100(20):11660–11665
 162. Magnani GS et al (2010) Diversity of endophytic bacteria in Brazilian sugarcane. *Genet Molec Res* 9(1):250–258
 163. Mendes R et al (2007) Diversity of cultivated endophytic bacteria from sugarcane: Genetic and biochemical characterization of *Burkholderia cepacia* complex isolates. *Appl Environ Microbiol* 73(22):7259–7267
 164. Editorial N (2010) Brazil's biotech boom. *Nature* 466(7304):295–295
 165. Bonfim K et al (2007) RNAi-mediated resistance to bean golden mosaic virus in genetically engineered common bean (*Phaseolus vulgaris*). *Molec Plant Microbe Interact* 20(6):717–726
 166. Oda L (2011) Approval of Brazilian transgenic beans has social importance, says ANBio in the Sacramento Bee. PRN News-wire, AnBio, Sao Paulo, p 2
 167. Paganelli A et al (2010) Glyphosate-based herbicides produce teratogenic effects on vertebrates by impairing retinoic acid signaling. *Chem Res Toxicol*
 168. Chassy B, Parrott W (2009) Is this study believable? Examples from animal studies with GM foods. *Agric Biotechnol* 9. doi: <http://www.agribiotech.info/details> and <http://www.botanischergarten.ch/Peer-Review/Chassy-Parrott-Believable-2009.doc>
 169. Antoniou M et al (2010) GM Soy, sustainable?, responsible?, GV-SOJA, Nachhaltig? Verantwortungsbewusst? German, p 11
 170. Martin-Orue SM et al (2002) Degradation of transgenic DNA from genetically modified soya and maize in human intestinal simulations. *Brit J Nutr* 87(6):533–542
 171. Netherwood T et al (2004) Assessing the survival of transgenic plant DNA in the human gastrointestinal tract. *Nat Biotechnol* 22(2):204–209
 172. Netherwood T et al (1999) Gene transfer in the gastrointestinal tract. *Appl Environ Microbiol* 65(11):5139–5141
 173. Zhang L et al (2011) Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Res*
 174. GMwatch (20110921) We incorporate genetic information from the food we eat – new study GMwatch website. <http://www.gmwatch.org>, DOI: http://www.gmwatch.org/index.php?option=com_content&view=article&id=13423:we-incorporate-genetic-information-from-the-food-we-eat-new-study
 175. Auer C, Frederick R (2009) Crop improvement using small RNAs: applications and predictive ecological risk assessments. *Trends Biotechnol* 27(11):644–651
 176. Fransen R (2007) Peer review: too much of a good thing? *Scientist* 21(9):18–18
 177. Waltz E (2009) Battlefield, papers suggesting that biotech crops might harm the environment attract a hail of abuse from other scientists, News feature. *Nature* 461:27–32
 178. Ammann K (2011) Review: is the impact of Bt maize on non-target insects significantly negative?. ASK-FORCE contribution AF-8 AF-8, 23, 20111002, DOI: <http://www.ask-force.org/web/AF-8-Lovei/AF-8-Lovei-Non-Target-20111002-opensource.pdf>
 179. Reiss T, Lacasa ID (2007) Benchmarking national biotechnology policy across Europe: a systems approach using quantitative and qualitative indicators. *Res Eval* 16(4):331–339

180. Linkov F, Lovalekar M, LaPorte R (2006) Scientific journals are "faith based": is there science behind peer review? *J Roy Soc Med* 99(12):596–598
181. Lubchenco J (1998) Entering the century of the environment: a new social contract for science. *Science* 279(5350):491–497
182. Linkov F, Lovalekar M, LaPorte R (2007) Quality control of epidemiological lectures online: scientific evaluation of peer review. *Croatian Med J* 48(2):249–255
183. Ammann K (2007) Evaluations faculty of 1000, manuscript and links. p 3
184. Hansen M (2009) Statement from Michael Hansen, CEO of Elsevier's health sciences division, regarding Australia based sponsored journal practices between 2000 and 2005. Elsevier Website 7 May 2009, doi: http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01203 and <http://www.ask-force.org/web/Peer-Review/Hansen-Statement-ELSEVIER-2009.pdf>
185. Goldacre B (2009) Peer review is flawed but the best we've got. In: *The Guardian*.
186. Smith R (2005) Medical journals are an extension of the marketing arm of pharmaceutical companies. *Plos Med* 2(5):364–366
187. Smith R (2003) Medical journals and pharmaceutical companies: uneasy bedfellows. *Brit Med J* 326(7400):1202–1205
188. Scott A (2007) Peer review and the relevance of science. *Futures* 39(7):827–845
189. Graff GD, Newcomb J (2003) Agricultural biotechnology at the crossroads, Part 1: the changing structure of the industry. *Bio-era*, p 26
190. Kostoff R (2002) Citation analysis of research performer quality. *Scientometrics* 53(1):49–71
191. Rosi-Marshall EJ et al (2007) Toxins in transgenic crop byproducts may affect headwater stream ecosystems. *Proc Nat Acad Sci USA* 104:16204–16208
192. Tank JL et al (2010) Occurrence of maize detritus and a transgenic insecticidal protein (Cry1Ab) within the stream network of an agricultural landscape. *Proc Nat Acad Sci*
193. Beachy RN et al (2008) The burden of proof: a response to Rosi-Marshall et al. *Proc Nat Acad Sci* 105:16204–16208
194. Parrott W (2008) Study of Bt impact on caddisflies overstates its conclusions: response to Rosi-Marshall et al. *Proc Nat Acad Sci* 105: E10
195. McHughen A et al (2007) Letter to the editor of PNAS, related to the publication of Rosi-Marshall, E. PNAS, Washington
196. Velimirov A et al (2008) Biological effects of transgenic maize NK603xMON810 fed in long term reproduction studies in mice, Report, in *Forschungsberichte der Sektion IV Band 3/2008*, Bundesministerium für Gesundheit Familie und Jugend Sektion IV (ed) Herausgeber, Medieninhaber und Hersteller: Bundesministerium für Gesundheit, Familie und Jugend, Sektion IV Radetzkystraße 2, 1031 Wien, p 109
197. Ammann K (20100407) Review: the Austrian experiment with mice fed with a hybrid GM maize from Monsanto, Part 1: background and Part 2: experiment. ASK-FORCE contribution AF-5 **AF-5**, Experiment: 19p and Background 8p doi: <http://www.ask-force.org/web/AF-5-Austrian-Micestudy/AF-5-Austrian-Experiment-20100407-opensource.pdf>, <http://www.ask-force.org/web/AF-5-Austrian-Micestudy/AF-5-Austrian-Experiment-20100407-web.doc>, <http://www.ask-force.org/web/AF-5-Austrian-Micestudy/AF-5-Austrian-Exp-Background-20090807-opensource.pdf>, <http://www.ask-force.org/web/AF-5-Austrian-Micestudy/AF-5-Austrian-Exp-Background-20090828-web.pdf>
198. Sinha G (2009) Up in arms. *Nat Biotechnol* 27(7):592–594
199. Dona A, Arvanitoyannis IS (2009) Health risks of genetically modified foods. *Crit Rev Food Sci Nutr* 49(2):164–175
200. Tang H, Tan S, Cheng X (2009) A survey on sentiment detection of reviews. *Exp Syst Appl* 36:10760–10773
201. Ammann K (20090828) Review: genomic misconception of transgenesis, The difference between GM- and non-GM-crops on the level of molecular processes has been overestimated. ASK-FORCE contribution AF-7 AF-7, 57 DOI: <http://www.ask-force.org/web/AF-7-Dona-rebuttal/AF-7-Dona-20090828-opensource.pdf>
202. Chassy BM (2009) Global regulation of transgenic crops. In: Kriz AL, Larkins BA (eds) *Molecular genetic approaches to maize improvement*. Springer, Berlin, pp 107–124
203. Intemann KK, de Melo-Martin I (2008) Regulating scientific research: should scientists be left alone? *Faseb J* 22(3): 654–658
204. de Melo-Martin I, Meghani Z (2008) Beyond risk – A more realistic risk - benefit analysis of agricultural biotechnologies. *Embo Reports* 9(4):302–306
205. Seralini GE, Cellier D, de Vendomois JS (2007) New analysis of a rat feeding study with a genetically modified maize reveals signs of hepatorenal toxicity. *Arch Environ Contamin Toxicol*:596–602
206. EFSA (2007) Safety and nutritional assessment of GM plants and derived food and feed: the role of animal feeding trials. *Food Chem Toxicol* 46(Suppl 1):S2–S70
207. EFSA (2007) Statement of the scientific panel on genetically modified organisms on the analysis of data from a 90-day rat feeding study with MON 863 maize. European Food Safety Authority, p 5
208. EFSA (2007) Press release: EFSA reaffirms its risk assessment of genetically modified maize MON 863 maize. European Food Safety Authority, p 5
209. Imposteurs. Tout (ou presque) sur le CRIIGEN 2011 (cited 12. Oct 2011. Available from: http://imposteurs.over-blog.com/pages/Tout_ou_presque_sur_le_CRIIGEN-4536267.html
210. Kuntz M (2011) Seralini critique: the latest opus of "parallel science" of Criigen from March 2011. Parallel Science (Website) 2011 (cited 12 Oct 2011). Available from: http://ddata.over-blog.com/xxxyyy/1/39/38/37/Critical_views_on_Seralini_20110710.pdf
211. Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171(4361):964–967
212. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. *Nature* 171(4356):737–738

213. Wilkins MHF et al (1953) Helical structure of crystalline deoxypentose nucleic acid. *Nature* 172(4382):759–762
214. Berg P et al (1975) Summary statement of asilomar conference on recombinant DNA-molecules. *Proc Nat Acad Sci USA* 72(6):1981–1984
215. Berg P, Singer M (1995) The recombinant-DNA controversy - 20 years later. *Bio-Technol* 13(10):1132–1134
216. Friedberg EC (2007) The writing life of James D. Watson. *Adler Museum Bull* 33(2):3–16
217. Klug A (2004) The discovery of the DNA double helix. *J Molec Biol* 335(1):3–26
218. Bennett D, Glasner P, Travis D (1986) The politics of uncertainty. Routledge and Kegan Paul plc, London, p 218
219. NRC (National-Research-Council) (1989) Field testing genetically modified organism. Framework for decisions. In: National Research Council (ed) Committee on scientific evaluation of the introduction of genetically modified microorganisms and plants into the environment, NAO Sciences. The National Academy Press, Washington, DC, p 184
220. Lehrman S (1992) Overregulation could damage united-states biotechnology, says report. *Nature* 359(6396):569–569
221. Mundell I (1992) Britain wrestles with EC rule on modified organisms. *Nature* 359(6396):569–569
222. McClintock B (1930) A cytological demonstration of the location of an interchange between two non-homologous chromosomes of *Zea mays*. *Proc Nat Acad Sci USA* 16: 791–796
223. McClintock B (1953) Induction of instability at selected loci in Maize. *Genetics* 38(6):579–599
224. Fedoroff N (1994) McClintock, Barbara (June 16, 1902 September 2, 1992). *Proc Am Philos Soc* 138(3):431–445
225. Fedoroff N, Schlappi M, Raina R (1995) Epigenetic regulation of the maize Spm transposon. *Bioessays* 17(4):291–297
226. Shapiro JA (1997) Genome organization, natural genetic engineering and adaptive mutation. *Trends Genet* 13(3): 98–104
227. Lewin R (1983) A naturalist of the genome. *Science* 222(4622):402–405
228. Arber W (2000) Genetic variation: molecular mechanisms and impact on microbial evolution. *Fems Microbiol Rev* 24(1):1–7
229. Arber W (2002) Roots, strategies and prospects of functional genomics. *Curr Sci* 83(7):826–828
230. Arber W (2003) Elements for a theory of molecular evolution. *Gene* 317(1–2):3–11
231. Arber W (2004) Biological evolution: lessons to be learned from microbial population biology and genetics. *Res Microbiol* 155(5):297–300
232. Arber W (1994) Molecular evolution: comparison of natural and engineered genetic variations. *Pontifical Acad Sci Scripta Varia* 103:90–101
233. Hackett P (2002) Genetic engineering: what are we fearing? *Trans Res* 11(2):97–99
234. Ghatnekar L, Jaarola M, Bengtsson BO (2006) The introgression of a functional nuclear gene from *Poa* to *Festuca ovina*. *Proc Biol Sci* 273(1585):395–399
235. Baudo MM et al (2006) Transgenesis has less impact on the transcriptome of wheat grain than conventional breeding. *Plant Biotechnol J* 4(4):369–380
236. Batista R et al (2008) Microarray analyses reveal that plant mutagenesis may induce more transcriptomic changes than transgene insertion. *Proc Nat Acad Sci USA* 105(9):3640–3645
237. Shewry PR et al (2007) Are GM and conventionally bred cereals really different? *Trends Food Sci Technol* 18(4): 201–209
238. Ammann K (2009) Feature: why farming with high tech methods should integrate elements of organic agriculture. *New Biotechnol* 25:378–388
239. Barnabás B, Obert B, Kovács G (1999) Colchicine, an efficient genome-doubling agent for maize (*Zea mays* L.) microspores cultured in anthero. *Plant Cell Reports* 18(10):858–862
240. Awoleye F et al (1994) Nuclear-DNA content and in-vitro induced somatic polyploidization cassava (*Manihot-Esculenta* crantz) breeding. *Euphytica* 76(3):195–202
241. Reynolds MP, van Ginkel M, Ribaut JM (2000) Avenues for genetic modification of radiation use efficiency in wheat. *J Exp Botany* 51:459–473
242. Molnar I, Benavente E, Molnar-Lang M (2009) Detection of intergenomic chromosome rearrangements in irradiated *Triticum aestivum* – *Aegilops biuncialis* amphiploids by multicolour genomic in situ hybridization. *Genome* 52(2):156–165
243. Schouten HJ, Jacobsen E (2007) Are mutations in genetically modified plants dangerous? *J Biomed Biotechnol*
244. Latham JR, Wilson AK, Steinbrecher RA (2006) The mutational consequences of plant transformation. *J Biomed Biotechnol* 2006:1–7
245. Wilson A, Latham J, Steinbrecher R (2006) Transformation-induced mutations in transgenic plants: analysis and bio-safety implications. *Biotechnol Genet Eng Rev* 23(11):1–26
246. Baarends WM, van der Laan R, Grootegeod JA (2001) DNA repair mechanisms and gametogenesis. *Reproduction* 121(1):31–39
247. Dong CM, Whitford R, Langridge P (2002) A DNA mismatch repair gene links to the Ph2 locus in wheat. *Genome* 45(1):116–124
248. Morikawa K, Shirakawa M (2001) Three-dimensional structural views of damaged-DNA recognition: T4 endonuclease V, E coli Vsr protein, and human nucleotide excision repair factor XPA (vol 460, pg 257, 2000). *Mutation Res DNA Repair* 485(3):267–268
249. Lammerts van Bueren ET, Struik PC, Jacobsen E (2002) Ecological concepts in organic farming and their consequences for an organic crop ideotype. *Netherlands J Agric Sci* 50(1):1–26
250. Lammerts van Bueren ET, Struik PC, Jacobsen E (2003) Organic propagation of seed and planting material: an overview of problems and challenges for research. *Njas-Wageningen J Sci* 51(3):263–277
251. Anonymous P (1992) Pose no special risks just because of the processes used to make them. *Nature* 356(6364):1–2

252. Andree P (2002) The biopolitics of genetically modified organisms in Canada. *J Canad Stud Revue D Etudes Canadiennes* 37(3):162–191
253. Berwald D, Carter CA, Gruere GP (2006) Rejecting new technology: the case of genetically modified wheat. *Am J Agric Econ* 88(2):432–447
254. Macdonald P, Yarrow S (2002) Regulation of Bt crops in Canada. In: 8th international colloquium on invertebrate pathology and microbial control/35th annual meeting of the SIP/6th international conference on *Bacillus Thuringiensis*. Academic Press Inc Elsevier Science, Iguassu Falls, Brazil
255. Ramjoue C (2007) The transatlantic rift in genetically modified food policy. *J Agric Environ Ethics* 20(5):419–436
256. Ramjoue C (2007) The transatlantic rift in genetically modified food policy. Thesis presented to the Faculty of Arts. University of Zurich, Zurich, p 263
257. Kalaitzandonakes N, Marks L, Vickner SS (2005) Sentiments and acts towards genetically modified foods. *Int J Biotechnol* 7(1–3):161–177
258. Snyder LU et al (2008) European union's moratorium impact on food biotechnology: a discussion-based scenario. *J Nat Resour Life Sci Educ* 37:27–31
259. Herman RA, Chassy BM, Parrott W (2009) Compositional assessment of transgenic crops: an idea whose time has passed. *Trends Biotechnol* 27(10):555–557, Corrected proof
260. Romeis J et al (2008) Assessment of risk of insect-resistant transgenic crops to nontarget arthropods. *Nat Biotechnol* 26(2):203–208
261. Raybould AF (2010) Reducing uncertainty in regulatory decision-making for transgenic crops: more ecological research or shrewder environmental risk assessment? *GM crops* 1(1):1–7
262. Raven P et al (2006) Where next for genome sequencing? *Science* 311(5760):468–468
263. Kesavan PC, Swaminathan MS (2008) Strategies and models for agricultural sustainability in developing Asian countries. *Philos Trans Roy Soc B-Biol Sci* 363:877–891
264. Plan D, van den Eede G (2010) The EU legislation on GMOs. JRC Scientific and Technical Reports. European Commission Joint Research Center, JRC, and Institute for Health and Consumer Protection IHCP. Publications Office of the European Union, © European Union, Luxembourg
265. McLean MA et al (2002) A conceptual framework for implementing biosafety: linking policy, capacity, and regulation. In: ISNAR briefing papers. ISNAR, International Service for National Agricultural Research, Washington DC, pp 1–12
266. Graff GD, Zilberman D, Bennett AB (2009) The contraction of agbiotech product quality innovation. *Nat Biotechnol* 27(8):702–704
267. Miller JK, Bradford KJ (2010) The regulatory bottleneck for biotech specialty crops. *Nat Biotechnol* 28(10):1012–1014
268. Strauss SH et al (2009) Strangled at birth? Forest biotech and the convention on biological diversity. *Nat Biotech* 27(6):519–527
269. McLean MA, Charest PJ (2000) The regulation of transgenic trees in North America. *Silvae Genetica* 49(6):233–239
270. Kalaitzandonakes N, Alston JM, Bradford KJ (2007) Compliance costs for regulatory approval of new biotech crops. *Nat Biotechnol* 25(5):509–511
271. Morandini P (2007) (20071211) A serious cover up story unveiled in Italy concerning a GM crop field trial, Press release. DOI: <http://www.botanischergarten.ch/ASK-FORCE-NEWS-Maize-Lombardia/Morandini-press-release-20071211.pdf>
272. Morandini P (2008) Al contadino non far sapere. *Espansione* n. 5–41, Polenta, May 2008, p 3
273. Marshall A (2007) Another inconvenient truth. In Europe, no one apparently wants to listen if you have good news about genetically modified organisms (GMOs). *Nat Biotechnol* 25(12):1330
274. Ammann K (2009) Biodiversity and GM crops. In: Ferry N, Gatehouse AMR (eds) *Environmental impact of genetically modified/novel crops*, released in March, 423p. CAB International, Wallingford, p 28
275. Ronald PC, Adamchak RW (2008) *Tomorrow's table: organic farming, genetics, and the future of food*. Oxford University Press, Oxford, p 232
276. deRenobales-Scheifler M (2009) More sustainable food: genetically modified seeds in organic farming. Junta General del Principado de Asturias Sociedad Internacional de Bioética (SIBI), Gijon, p 119
277. Paarlberg R (2009) The ethics of modern agriculture. *Society* 46(1):4–8
278. Herring RJ (2007) The genomics revolution and development studies: science, poverty and politics. *J Develop Stud* 43(1):1–30
279. Paarlberg RL (2002) The real threat to GM crops in poor countries: consumer and policy resistance to GM foods in rich countries. *Food Policy* 27(3):247–250
280. Driessen PL (2006) *Eco-imperialism: green power–black death*. Academic Foundation, New Delhi
281. Alene AD, Coulibaly O (2009) The impact of agricultural research on productivity and poverty in sub-Saharan Africa. *Food Policy* 34(2):198–209
282. Gruère G, Sengupta D (2009) GM-free private standards and their effects on biosafety decision-making in developing countries. *Food Policy* 34(5):399–406
283. Spielman DJ, Cohen JI, Zambrano P (2007) Are developing-country policies and investments promoting research and research partnerships in agricultural biotechnology? *Int J Biotechnol* 9(6): ISSN 0963-6048(print)|1741-5020(electronic)
284. Cohen JI (2005) Poorer nations turn to publicly developed GM crops. *Nat Biotechnol* 23(1):27–33
285. Alhassan WS (2002) Agrobiotechnology application in West and Central Africa (2002 Survey outcome). CORAF/WECARD–IITA International Institute of Tropical Agriculture, Ibadan, p 107
286. Gruere G, Bouët A, Mevel S (2007) Genetically modified food and international trade: the case of India, Bangladesh,

- Indonesia and the Philippines. In: IFPRI Discussion Paper 00740. IFPRI, Washington, p 60
287. Smale M et al (2008) The economic impact of transgenic crops in developing countries: a note on the methods. *Int J Biotechnol* 10(6):519–555
 288. Falck-Zepeda JB, Traxler G, Nelson RG (2000) Surplus distribution from the introduction of a biotechnology innovation. *Am J Agric Econ* 82(2):360–369
 289. Pray CE et al (2006) Costs and enforcement of biosafety regulations in India and China. *Int J Technol Globalisation* 2(1–2):137–57
 290. Antle JM (1999) Benefits and costs of food safety regulation. *Food Policy* 24(6):605–623
 291. Shelton AM (2003) Considerations for conducting research in agricultural biotechnology. *J Invertebr Pathol* 83(2):110–112
 292. Kochetkova T (2006) The transatlantic conflict over GM food: cultural background. In: Kaiser M, Lien M (eds) *Ethics and the politics of food*. Wageningen Academic, Wageningen, pp 325–329
 293. Laget P, Cantley M (2001) European responses to biotechnology: research, regulation, and dialogue. *Issues Sci Technol* 17(4):37–42
 294. Ramessar K et al (2010) Going to ridiculous lengths (mdash) European coexistence regulations for GM crops. *Nat Biotech* 28(2):133–136
 295. Bradford KJ et al (2005) Regulating transgenic crops sensibly: lessons from plant breeding, biotechnology and genomics. *Nat Biotechnol* 23(4):439–444
 296. Stein AJ et al (2007) Plant breeding to control zinc deficiency in India: how cost-effective is biofortification? *Public Health Nutr* 10(5):492–501
 297. Stein AJ, Qaim M (2007) The human and economic cost of hidden hunger. *Food Nutr Bull* 28(2):125–134
 298. Stein AJ, Sachdev HPS, Qaim M (2007) What we know and don't know about golden rice. *Nat Biotechnol* 25(6):624–624
 299. Humphrey JH et al (1998) Neonatal vitamin A supplementation: effect on development and growth at 3 y of age. *Am J Clin Nutr* 68(1):109–117
 300. Humphrey JH, West KP, Sommer A (1992) Vitamin-a deficiency and attributable mortality among under-5-year-olds. *Bull World Health Organization* 70(2):225–232
 301. Depee S et al (1995) Lack of improvement in vitamin-a status with increased consumption of dark-green leafy vegetables. *Lancet* 346(8967):75–81
 302. Mayer JE, Pfeiffer WH, Beyer P (2008) Biofortified crops to alleviate micronutrient malnutrition. *Genome studies Molec Genet* edited by Juliette de Meaux and Maarten Koornneef/ *Plant Biotechnol*, edited by Andy Greenland and Jan Leach 11(2):166–170
 303. Miller HI (2009) A golden opportunity, squandered. *Trends Biotechnol* 27(3):129–130
 304. Potrykus I (2003) Nutritionally enhanced rice to combat malnutrition disorders of the poor. *Nutr Rev* 61(6):S101–S104
 305. Stein AJ et al (2008) Potential impacts of iron biofortification in India. *Soc Sci Med* 66(8):1797–1808
 306. Stein AJ, Sachdev HPS, Qaim M (2006) Potential impact and cost-effectiveness of Golden Rice. *Nat Biotechnol* 24(10):1200–1201
 307. Qaim M, Stein AJ (2008) Economic consequences of Golden Rice. In: Invited presentation at the fourth conference of the European plant science organisation. Toulon (Cote d'Azur), France
 308. Qaim M, Stein AJ, Meenakshi JV (2007) Economics of biofortification. In: Otsuka K, Kalirajan K (eds) *Contributions of agricultural economics to critical policy issues*. Blackwell, Malden, pp 119–133
 309. Qaim M, Pray CE, Zilberman D (2008) Economic and social considerations in the adoption of Bt crops. In: Romeis J, Shelton AM, Kennedy GG (eds) *Integration of insect-resistant genetically modified crops within IPM programs*. Springer, Dordrecht, pp 329–356
 310. Bouis HE (2007) The potential of genetically modified food crops to improve human nutrition in developing countries. *J Develop Stud* 43(1):79–96
 311. Atanassov AB, Brink A, Burachik J, Cohen M, Dhawan JI, Eboru V, Falck-Zepeda RV, Herrera-Estrella J, Komen L, Low J, Omaliko FC, Odhiambo E, Quemada B, Peng H, Sampaio Y, Sithole-Niang MJ, Sittenfeld I, Smale A, Sutrisno M, Valyasevi R, Zafar Y, Zambrano P (2004) To reach the poor: results from the ISNAR-IFPRI next harvest study on genetically modified crops, in EPTD Discussion Paper No. 116. 2004, ISNAR-IFPRI, International Food Policy Research Institute, Washington DC
 312. Potrykus I (2010) Regulation must be revolutionized. *Nature* 466(7306):561–561
 313. Dhlamini Z et al (2005) Status of research and application of crop technologies in developing countries, preliminary assessment. In: FAO (ed) *FAO Reports*, FAO, Rome, p 62
 314. Krattiger A, Mahoney RT (2006) Intellectual property and public health. *Bull World Health Organization* 84(5):340–340
 315. Atkinson RC et al (2003) Public sector collaboration for agricultural IP management including corrigendum of fig. on front page of text, vol. 302, 5648, pp 1152–1152. *Science* 301(5630):174–175
 316. Beachy R et al (2002) Divergent perspectives on GM food. *Nat Biotechnol* 20(12):1195–1196
 317. Krattiger A, Mahoney RTL, Nelsen L, Thompson GA, Bennett AB, Satyanarayana K, Graff GD, Fernandez C, Kowalsky SP (2007) *Intellectual property management in health and agricultural innovation a handbook of best practice*. MIHR/PIPR, Oxford/Davis, pp 1539–1559
 318. Singh A, Hallihsor S, Rangan L (2009) Changing landscape in biotechnology patenting. *World Patent Information*, pp 219–225
 319. Wright B (2008) Plant genetic engineering and intellectual property protection. *Agricultural Biotechnology in California Series Publication*, no. 8186
 320. Lawson C (2004) Patents and the CGIAR system of international agricultural research centres' germplasm collections under the International Treaty on Plant Genetic resources for food and agriculture. *Aust J Agric Res* 55(3):307–313

321. Delmer DP et al (2003) Intellectual property resources for international development in agriculture. *Plant Physiol* 133(4):1666–1670
322. Lempert DH (2009) A dependency in development indicator for NGOs and international organizations. *Global Jurist* 9(2): Article 6
323. Neidecker-Gonzales O, Nestel P, Bouis H (2007) Estimating the global costs of vitamin A capsule supplementation: a review of the literature. *Food Nutr Bull* 28:307–316
324. Gressel J, Zilberstein A (2003) Let them eat (GM) straw. *Trends Biotechnol* 21(12):525–530
325. Potrykus I (2010) Constraints to biotechnology introduction for poverty alleviation. *New Biotechnol* 27(5):447–448
326. Ademola AA (2011) Global capture of crop biotechnology in developing world over a decade. *J Genet Eng Biotechnol* (in press)
327. Taverne D (2007) *The March of unreason*. Oxford University Press, Oxford, p 320
328. Durant J (2005) The march of unreason: science, democracy, and the new fundamentalism. *Nature* 435(7040):277–278
329. Taverne D (2005) The new fundamentalism, Commentary. *Nat Biotechnol* 23(4):415–416
330. Herring RJ (2008) Whose numbers count? Probing discrepant evidence on transgenic cotton in the Warangal district of India. *Int J Mult Res Approach* 2:145–159
331. Marris E (2006) Environmental activism: in the name of nature. *Nature* 443(7111):498–501
332. Atkinson HJ, Urwin PE (2008) Europe needs to protect its transgenic crop research. *Nature* 453(7198):979–979
333. Leader SH, Probst P (2003) The earth liberation front and environmental terrorism. *Terrorism Polit Violence* 15(4):37–58
334. Finkel E (2011) Vandals attack transgenic wheat test plot. *Science Insider*, July 2011
335. Bettles C (2011) Scientist distances himself from activists. *Farm online*. DOI: <http://sl.farmonline.com.au/news/nationalrural/grains-andcropping/cereal/scientist-distances-himself-from-activists/2239218.aspx?storypage=0> and <http://www.ask-force.org/web/Field-Destruction/Bettles-Scientist-Distances-Schubert-20110728.pdf>
336. Kuntz M (2011) Academic and governmental research on GMOs has been the target of numerous acts of vandalism in Europe. OGM, environnement, santé et politique. DOI: <http://www.marcel-kuntz-ogm.fr/article-news-55055856.html>, news in English, French and Spanish and <http://ddata.over-blog.com/xxxyyy/1/39/38/37/public-research-vandalized.pdf> and <http://www.marcel-kuntz-ogm.fr/article-news-55055856.html> and <http://www.ask-force.org/web/Field-Destruction/Kuntz-Public-Government-Research-Vandalism-Europe-2011.pdf>
337. Da Silva W (2011) In focus: the sad, sad demise of Greenpeace in cosmos. About Luna Media Pty Ltd, the boutique publishing company behind COSMOS. Sidney, Australia
338. Gough M (2011) Greenpeace destroys CSIRO wheat GM trial in Cosmos. About Luna Media Pty Ltd, the boutique publishing company behind COSMOS. Sidney, Australia
339. Smith J (2003) *Seeds of deception*. Yes! Books, Iowa, p 304
340. Smith J (2007) *Genetic roulette, the documented health risks of genetically engineered foods*. YES ! Books and Chelsea Green, Fairfield Iowa, p 319, second printing edn
341. Miller H (2008) Auf wiedersehen, academic freedom. *Wall Street J Europe*. p 3
342. Rao CK (2010) *Moratorium on Bt Brinjal, a review of the order of the Minister of Environment and Forests, Government of India*. Foundation for biotechnology awareness and education, Bangalore, p 74
343. Weese TL, Bohs L (2010) Eggplant origins: out of Africa, into the orient. *Taxon* 59(1):49–56
344. PRRI (2006) Correspondence between PRRI (Public Research and Regulation Initiative) and FoE (Friends of the Earth). www.pubresreg.org. DOI: <http://www.ask-force.org/web/PRRI-FoE/PRRI-FoE-Corresp-Letter-to-FoE-20060629.pdf> and <http://www.ask-force.org/web/PRRI-FoE/PRRI-FoE-Corresp-Answer-to-FoE-20060926.pdf> and <http://www.ask-force.org/web/PRRI-FoE/PRRI-FoE-Corresp-Answer-FoE-to-PRRI-20060703.pdf>
345. Apel A (2010) The costly benefits of opposing agricultural biotechnology. *New Biotechnol* 27(5):635–640
346. Borlaug NE (2000) Ending world hunger. The promise of biotechnology and the threat of antiscience zealotry. *Plant Physiol* 124(2):487–490
347. Hemming D (2006) Swiss vote encourages Austria's anti-GM stance. *Outlook Agric* 35(1):82–82
348. Motion J, Weaver CK (2005) The epistemic struggle for credibility: rethinking media relations. *J Commun Manag* 9(3):246–255
349. Burke D (2004) GM food and crops: what went wrong in the UK? Many of the public's concerns have little to do with science. *Embo Reports* 5(5):432–436
350. Blas X (2009) Bill Gates shifts focus to fighting hunger. *Financial Times*, London, p 1
351. Miller H, Morandini P, Ammann K (2008) Is biotechnology a victim of anti-science bias in scientific journals? *Trends Biotechnol* 26(3):122–125, Electronic Prepublication 17 Feb 2008, Hardcopy available in March
352. Horton R (1999) Secret society – Scientific peer review and Pusztai's potatoes. *Tls-the Times Literary Suppl* 5046:8–9
353. Horton R (1999) GM food debate – Editors reply. *The Lancet* 354(9191):1729–1729
354. Horton R (1999) Genetically modified foods: "absurd" concern or welcome dialogue? *The Lancet* 354(9187):1314–1315
355. Horton R (1999) Health risks of genetically modified foods, editorial, reply to Mitchell and Bradbury, *Lancet*, p. 1769. *The Lancet* 353(9167):1811–1811
356. Horton R (1999) Scientific misconduct: exaggerated fear but still real and requiring a proportionate response. *The Lancet* 354(9172):7–8
357. Ammann K (20110111) Review: Arpad Pusztai's feeding experiments of GM potatoes with lectins to rats: anatomy of a controversy 1998–2009. ASK-FORCE contribution AF-2 AF-2, 46. DOI: <http://www.ask-force.org/web/AF-2-Pusztai/AF-2-Pusztai-Food-Safety-20110111.opensource.pdf>

358. Quist D, Chapela IH (2001) Transgenic DNA introgressed into traditional maize landraces in Oaxaca, Mexico. *Nature* 414(6863):541–543
359. Campbell P (2002) Quist-chapela paper: editorial note 2. *Nature* 417(6892):897–897, 27 June 2002
360. Pineyro-Nelson A et al (2009) Transgenes in Mexican maize: molecular evidence and methodological considerations for GMO detection in landrace populations. *Molec Ecol* 18(4): 750–761
361. Schubert D, Tribe D. comments (2006) Three faces of science fraud. GMO Pundit, DOI: <http://gmopundit.blogspot.com/2006/02/david-schubert-alleges-systematic.html>.
362. Bradford KJ et al (2005) Regulatory regimes for transgenic crops – Response. *Nat Biotechnol* 23(7):787–789
363. Schubert D (2005) Regulatory regimes for transgenic crops. *Nat Biotechnol* 23(7):785–787
364. Punnett RC (1928) Scientific papers of William Bateson. Cambridge University Press, Cambridge
365. Strick J (1999) Darwinism and the origin of life: the role of H.C. Bastian in the British spontaneous generation debates, 1868–1873. *J History Biol* 32(1):51–92
366. Chassy BN (2002) Food safety evaluation of crops produced through biotechnology. *J Am Coll Nutr* 21(3):166S–173S
367. Chassy B et al (2007) Nutritional and safety assessments of foods and feeds nutritionally improved through biotechnology: case studies. *J Food Sci* 72:R131–R137
368. Shelton AM et al (2009) Appropriate analytical methods are necessary to assess nontarget effects of insecticidal proteins in GM crops through meta-analysis (Response to Andow et al. 2009). *Environ Entomol* 38(6):1533–1538
369. Duan JJ et al (2010) Extrapolating non-target risk of Bt crops from laboratory to field. *Biol Lett* 6(1):74–77
370. Broer I et al (2011) Response to the criticism by Taube et al. in ESE 23:1, 2011, on the booklet “Green Genetic Engineering”. German Research Foundation (DFG), Environmental Sciences Europe, vol 23, issue 1, p 16
371. Green JM, Owen MDK (2010) Herbicide-resistant crops: utilities and limitations for herbicide-resistant weed management *J Agric Food Chem*
372. Johnson WG et al (2009) Influence of glyphosate-resistant cropping systems on weed species shifts and glyphosate-resistant weed populations. *Euro J Agron* 31(3):162–172
373. Duke SO, Powles S (2009) Glyphosate-resistant crops and weeds: now and in the future. *AgBioForum* 12(3&4):346–357
374. Vila-Aiub MM et al (2008) Glyphosate-resistant weeds of South American cropping systems: an overview. *Pest Manag Sci* 64(4):366–371
375. Powles SB (2008) Evolved glyphosate-resistant weeds around the world: lessons to be learnt. *Pest Manag Sci* 64(4):360–365
376. Powles SB, Preston C (2006) Evolved glyphosate resistance in plants: biochemical and genetic basis of resistance. *Weed Technol* 20(2):282–289
377. Neve P (2008) Simulation modelling to understand the evolution and management of glyphosate resistant in weeds. *Pest Manag Sci* 64(4):392–401
378. Mikulka J, Chodova D (2000) Long-term study on the occurrence of weeds resistant to herbicides in the Czech Republic. (*Zeitschrift Fur Pflanzenkrankheiten Und Pflanzenschutz*) *J Plant Dis Protect* 107:373–376
379. Hilbeck A, Schmidt JEU (2006) Another view on Bt proteins – how specific are they and what else might they do? *Biopest Int* 2(1):1–50
380. Hilbeck A, Meier M, Raps A (2000) Review on non-target organisms and Bt-plants. *Ecostrat GmbH, Ecological Technology Assessment Consulting, Amsterdam*, p 80
381. Hilbeck A et al (1998) Toxicity of *Bacillus thuringiensis* Cry1Ab toxin to the predator *Chrysoperla carnea* (Neuroptera: Chrysopidae). *Environ Entomol* 27(5):1255–1263
382. Hilbeck A et al (1999) Prey-mediated effects of Cry1Ab toxin and protoxin and Cry2A protoxin on the predator *Chrysoperla carnea*. *Entomol Exper Et Applicata* 91(2): 305–316
383. Romeis J, Dutton A, Bigler F (2004) *Bacillus thuringiensis* toxin (Cry1Ab) has no direct effect on larvae of the green lacewing *Chrysoperla carnea* (Stephens) (Neuroptera: Chrysopidae). *J Insect Physiol* 50(2–3):175–183
384. Marshall A (2007) GM soybeans and health safety – a controversy reexamined, additional texts. *Nat Biotechnol* 25(9):981–987
385. Ammann K (2009) Review web version: are rat organs damaged after feeding on GM soybeans? The Ermakova Case. ASK-FORCE contribution No. 4, 20, 20090801. DOI: <http://www.ask-force.org/web/AF-4-Ermakova/AF-4-Ermakova-20090828-web.pdf>
386. Seralini GE, Cellier D, de Vendomois JS (2007) New analysis of a rat feeding study with a genetically modified maize reveals signs of hepatorenal toxicity. *Arch Environ Contamin Toxicol* 52:596–602
387. OECD (1998) 408 repeated dose 90-day oral toxicity study in rodents (Updated Guideline, Adopted 21st Sept 1998)
388. OECD (1998) 407 repeated dose 28-day oral toxicity study in rodents, Adopted by the Council on 27th July 1995, in OECD Guideline for the testing of chemicals
389. Ammann K (20110921) Summary of 11 ASK-FORCE contributions on biosafety of biotechnology crops. ASK-FORCE contributions AF summary, 69. DOI: <http://www.ask-force.org/web/ASK-FORCE-Summary/ASK-FORCE-Summary.pdf>
390. Candolfi MP et al (2004) A faunistic approach to assess potential side-effects of genetically modified Bt-corn on non-target arthropods under field conditions. *Biocontrol Sci Technol* 14(2):129–170
391. Marvier M et al (2007) A meta-analysis of effects of Bt cotton and maize on nontarget invertebrates. *Science* 316(5830):1475–1477. doi:10.1126/science.1139208
392. Wolfenbarger LL et al (2008) Bt crop effects on functional guilds of non-target arthropods: a meta-analysis. *PLoS ONE* 3(5):e2118
393. Naranjo SE (2009) Impacts of Bt crops on non-target invertebrates and insecticide use patterns. *CAB Rev Perspect Agric Veterinary Sci, Nutr Nat Resour* 4:11–23

394. Duan JJ et al (2008) A meta-analysis of effects of Bt crops on honey bees (Hymenoptera: Apidae). *PLoS ONE* 3(1):e1415
395. Ammann Ki et al (2004) Biosafety in agriculture: is it justified to compare directly with natural habitats? *Frontiers in Ecology, Forum: GM crops: balancing predictions of promise and peril*, vol 2, pp 54–160
396. Ammann K (2005) Effects of biotechnology on biodiversity: herbicide-tolerant and insect-resistant GM crops. *Trends Biotechnol* 23(8):388–394
397. PRRI Public Research and Regulation Initiative(2009) Letter to CBD: LMOs that are likely to have adverse environmental impacts. 20090914, Downloads of PRRI www.pubresreg.org, 3 DOI: http://www.pubresreg.org/index.php?option=com_docman&task=doc_download&gid=490
398. Gupta A (2010) Transparency to what end? Governing by disclosure through the biosafety clearing house. *Environ Plann C-Govern Policy* 28(1):128–144
399. Felke M et al (2010) Effect of Bt-176 maize pollen on first instar larvae of the Peacock butterfly (*Inachis io*) (Lepidoptera; Nymphalidae). *Environ Biosafety Res* 9(1):5–12, Received: 20 November 2008, Accepted: 5 December 2009, online publication 28. October 2010
400. Ammann K (20090911) ASK-FORCE structure and possible contributions. ASK-FORCE contributions, 12 DOI: <http://www.botanischergarten.ch/ASK-FORCE-Strategy/ASK-FORCE-General-List-20090911.pdf>
401. Freeland C A hard-pressed trade. In: *Financial times*, 20070504. Financial Times, London
402. Moore A (2006) Bad science in the headlines – Who takes responsibility when science is distorted in the mass media? *Embo Reports* 7(12):1193–1196
403. Erjavec K, Erjavec E (2009) Changing EU agricultural policy discourses? The discourse analysis of Commissioner's speeches 2000–2007. *Food Policy* 34(2):218–226
404. Brabeck-Lemathe P (2008) Nestlé Chairman calls on European policymakers to reconsider opposition to genetically modified (GM) crops; Says 10,000 liters of water required to produce as little as one to two liters of biodiesel. In: *Finfacts Business and Finance Portal*. Finfacts, Irland, 6 pp. <http://www.ask-force.org/web/Nestle/Brabeck-Finfact-Ireland-Interview-2009.pdf>
405. Rogers C, Farson RE (2007) Active listening, excerpt 1957. University of Chicago Industrial Relations Center, Gordon Training International, Chicago
406. Conklin J (2005) Wicked problems and social complexity. In: Conklin J (ed) *Dialogue mapping: building shared understanding of wicked problems*. Wiley, Chichester, p 20
407. Peter LJ, Hull R (2009) *The Peter principle, why things always go wrong*. HarperCollins, New York, 192 pp
408. Blackmore C (2007) What kinds of knowledge, knowing and learning are required for addressing resource dilemmas?: a theoretical overview. *Environ Sci Policy* 10(6):512–525
409. Zwart NH (2007) Genomics and self-knowledge: implications for societal research and debate. *New Genet Soc* 26(2):181–202
410. Ikerd JE (1993) The need for a system approach to sustainable agriculture. *Agric Ecosyst Environ* 46(1–4):147–160
411. Fairclough N (2009) Language and globalization. *Semiotica* 173(1–4):317–342
412. Scoones I (2008) Mobilizing against GM crops in India, South Africa and Brazil. *J Agrarian Change* 8(2–3):315–344
413. Weissmann G (2006) DDT is back: let us spray! *Faseb J* 20:2427–2429
414. WHO (2005) WHO position on DDT use. In: WHO (ed) *Disease vector control under the stockholm convention on persistent organic pollutants*. WHO, Geneva, p 2
415. Tren R, Bate R (2001) Malaria and the DDT story. In: T.I.o.E.A (ed) *The Institute of Economic Affairs*, London, 112 pp
416. Losey JE, Raynor LS, Carter ME (1999) Transgenic pollen harms Monarch larvae. *Nature* 399:214
417. Ammann K (2007) Reconciling traditional knowledge with modern agriculture: a guide for building bridges. In: Krattiger A, Mahoney RTL, Nelsen L, Thompson GA, Bennett AB, Satyanarayana K, Graff GD, Fernandez C, Kowalsky SP (eds) *Intellectual property management in health and agricultural innovation a handbook of best practices*, Chapter 16.7. MIHR, PIPRA, Oxford/Davis, pp 1539–1559
418. Gerhards J (1997) The discursive versus the liberal public sphere: an empirical critique of Jurgen Habermas' concept of the public sphere. *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie* 49(1):1–34
419. Conklin J (2003) Wicked problems and fragmentation. (cited 2003; White Papers, This paper is Chapter 2 in *Dialog mapping: making sense of project fragmentation* Conklin J, forthcoming). Available from: <http://www.cognexus.org/id29.htm>
420. Fischer G (2000) Symmetry of ignorance, social creativity, and meta-design. *Knowledge-Based Syst* 13(7–8):527–537
421. Ammann K (2004) The role of science in the application of the precautionary approach. In: Fischer R, Schillberg S (eds) *Molecular farming, Plant-made Pharmaceuticals and Technical Proteins*. Wiley-VCH Verlag GmbH & Co KGaA, Weinheim, pp 291–302
422. Rittel H (1984) Second generation design methods. In: Cross N (ed) *Developments in design methodology*. Wiley, New York, pp 317–327
423. Rith C, Dubberly H (2007) Why Horst W. J. Rittel Matters. *Des Issues* 23(1):72–74
424. Rith C et al (2007) Bibliography of Horst W.J. Rittel. *Des Issues* 23(1):78–88
425. Schmidt I et al (2004) SEEBalance reg – managing sustainability of products and processes with the socio-eco-efficiency analysis by BASF. *Greener Manag Int* 45:79–94
426. Renn O (2008) *Risk Governance, coping with uncertainty in a complex world*. Earthscan, London
427. Moirand S (2003) Communicative and cognitive dimensions of discourse on science in the French mass media. *Discourse Stud* 5(2):175–206
428. Chiapello E, Fairclough N (2002) Understanding the new management ideology: a transdisciplinary contribution from critical discourse analysis and new sociology of capitalism. *Discourse Soc* 13(2):185–208
429. Motion J, Leitch S (1996) A discursive perspective from New Zealand: another world view. *Public Relat Rev* 22(3):297–309

430. Clark CE (2000) Differences between public relations and corporate social responsibility: an analysis. *Public Relat Rev* 26(3):363–380
431. Galtung J, Ruge MH (1965) The structure of foreign-news – the presentation of the congo, cuba and cyprus crises in 4 norwegian newspapers. *J Peace Res* 2(1):64–91
432. Renn O (2006) Risk communication – consumers between information and irritation. *J Risk Res* 9(8):833–849
433. Chen GKC (1975) What is systems-approach. *Interfaces* 6(1):32–37
434. Priest SH, Bonfadelli H, Rusanen M (2003) The “trust gap” hypothesis: predicting support for biotechnology across national cultures as a function of trust in actors. *Risk Anal* 23(4):751–766
435. Iyengar S et al (2009) “Dark areas of ignorance” revisited comparing international affairs knowledge in Switzerland and the United States. *Commun Res* 36(3):341–358
436. Bonfadelli H, Dahinden U, Leonarz M (2002) Biotechnology in Switzerland: high on the public agenda, but only moderate support. *Public Understand Sci* 11(2):113–130
437. von Grebmer K, Omamo SW (2007) Options for a rational dialogue on the acceptance of biotechnology. *Biotechnol J* 2(9):1121–1128
438. Huang JC, Newell S (2003) Knowledge integration processes and dynamics within the context of cross-functional projects. *Int J Project Manag* 21(3):167–176
439. Beer S (2004) Reflections of a cybernetician on the practice of planning. *Kybernetes* 33(3–4):767–773
440. Feldman M, Lowe N (2008) Consensus from controversy: Cambridge’s biosafety ordinance and the anchoring of the biotech industry. *Euro Plann Stud* 16(3):395–410
441. Bogner A (2010) Participation as a laboratory experiment paradoxes of deliberation on technology issues by lay people. *Zeitschrift Fur Soziologie* 39(2):87–105
442. Moore P (2000) Trees are the answer. *Forest Prod J* 50(10):12–19
443. Moore P (2000) A challenge: protect biodiversity and produce wood. *J Forestry* 98(8):A2–A3
444. Moore P (2002) Communication through participation, getting it right, environmentalism for the 21st Century. In: Ammann K, Papazova AB (eds) 1st dialogue on science. *Academia Engelberg, Engelberg*
445. Kahane A (2004) Solving Tough Problems: An Open Way of Talking, Listening, and Creating New Realities. In: Baetz BW (ed) Berrett-Koehler Publishers, San Francisco, 150 pp
446. Schenkel R (2010) The challenge of feeding scientific advice into policy-making. *Science* 330(6012):1749–1751
447. Rogers-Hayden T, Campbell JR (2003) Re-negotiating science in environmentalists’ submissions to New Zealand’s royal commission on genetic modification. *Environ Values* 12:515–534
448. Reich KH (2008) Science-and-religion/spirituality/theology dialogue: what for and by whom? *Zygon* 43(3):705–718
449. Papazova AB (2010) What do we need as visionaries: progress or development? abstract biovision 2010. *Biovision 2010*, DOI: <http://www.bibalex.org/bva2010/speakers/SpeakerDetails.aspx?m=1&sp=XOHvrH47wZRXTP5lzyEvA>

Grain Quality in Oil and Cereal Crops

DÉBORAH P. RONDANINI^{1,2}, LUCAS BORRÁS^{2,3},
ROXANA SAVIN⁴

¹Department of Crop Production, University of Buenos Aires, Buenos Aires, Argentina

²CONICET, National Council of Scientific and Technical Research, Buenos Aires, Argentina

³Departamento de Producción Vegetal, Universidad Nacional de Rosario, Zavalla, Santa Fe, Argentina

⁴Department of Crop and Forest Sciences, University of Lleida, Lleida, Spain

Article Outline

Glossary
Definition of the Subject
Grain Quality: Concept and Importance
Grain Structure
Grain Growth and Source–Sink Balance
Synthesis of Major Components
Main Factors Affecting Grain Quality
Future Directions
Bibliography

Glossary

Cereals Monocotyledon plant grains that accumulate starch as the main storage substance for subsequent germination. Two types have been distinguished – cereals that contain gluten and are used for bread-making (wheat, oats, barley, rye) and cereals that do not contain gluten (rice, maize).

Genotype × environment interaction Relative changes in genotype performance when grown under different environments.

Grain development Structural and functional changes that occur in the fertilized flower producing a mature grain capable of germinating.

Grain growth Irreversible increase in grain weight and size caused by cell division, expansion, and reserves accumulation.

Grain quality Group of grain characteristics and measurable attributes (objectively or subjectively) to meet the clients’ requirements (i.e., customer, industry, consumers).

Oilseeds Dicotyledon plant grains that accumulate oil as the main storage substance for subsequent germination. Oilseed crop seeds (sunflower, rapeseed, ground pea) are composed of 40–50% oil and 20–30% protein while proteo-oil crop seeds (soybean, lupine) comprise 15–30% oil and 30–40% protein.

Photoassimilates Carbohydrates (sugars, starch, or fructans, depending on the species) synthesized by the green plant parts and translocated to actively growing organs, like grains. Photoassimilates may originate from current photosynthesis or reserve remobilization.

Source–sink balance Quantitative relationship between plant photosynthetic capacity (source) and number of organs under active growth (sink) that are sustained by the former.

Plant stress Changes in plant metabolism in response to environments that endanger plant survival or hinder reaching maximum reproductive capacity.

Definition of the Subject

Grain quality is frequently regarded by agronomists and breeders to be as important as yield. Quality characteristics are the reason why only few plant species are used to satisfy most human requirements for food and fiber [1]. Grain quality comprises a group of characteristics that collectively determine the usefulness of the harvested grains for a particular end use. Therefore, to breed and manage grain crops to achieve a specific quality standard and to be able to predict the quality of a particular crop in a particular growing environment is rather important. Achieving this objective is dependent upon the knowledge of the factors modifying grain composition, and consequently grain quality.

As grain markets have become more specialized, there is a growing pressure on farmers to produce grains with greater uniformity and with certain characteristics [2]. Appropriate husbandry to obtain grains with high and stable “quality” will likely be of increasing importance in achieving economic benefits. It is well known that grain quality is modified by the environment and the crop management practices used by farmers. However, the strategies and tools required to produce grains with certain quality characteristics are not as well established as the ones for achieving high

yields. In this context, improving the understanding of the factors that determine grain quality has become increasingly important.

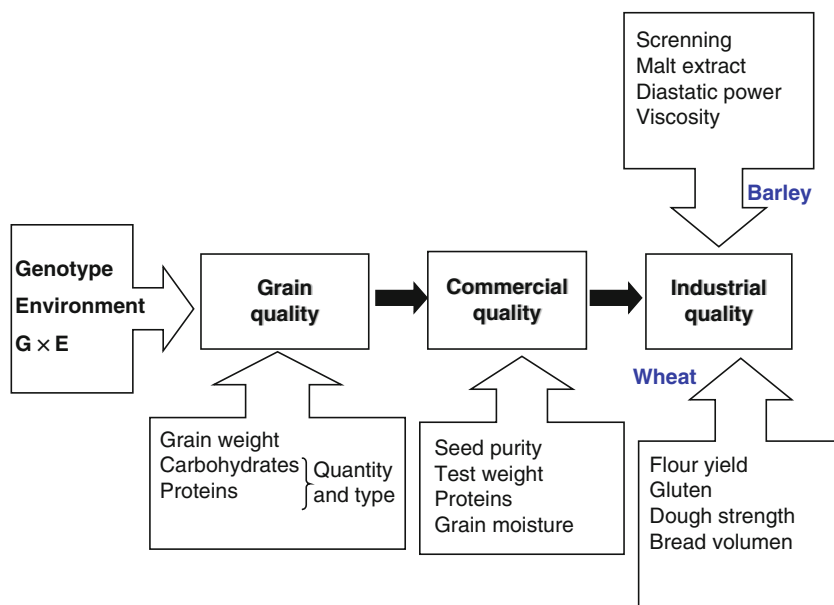
Grain Quality: Concept and Importance

In field crops, the quality of the end product is related to the composition and structure of the seed at harvest maturity. Seed composition and structure at harvest are determined by the genotype, the environment, and the crop management practices used during the crop growing cycle. It is not possible to propose a unique grain quality definition for any specie because it changes depending on the product end use. There is a proper criterion on the concept of quality for each specific end use and for each stage of the commercial chain in every crop (i.e., from harvest in the field, through grain dealers to the industry, Fig. 1). In this context, quality will be considered in relation to the criteria used by those involved in the various aspects of growth and utilization of the grain. As an example, for wheat and barley (Fig. 1), grain quality at the moment of harvest in the field is related to grain size (and weight) and the carbohydrates and protein composition. When the grain is sold to the grain dealer, seed purity, test weight, grain moisture, and protein percentage are the main characteristics that are taken into account for the prize (Fig. 1). After this stage, other attributes may be relevant and they will depend on the involved industry. For baking industry, flour yield and dough strength will be of maximum importance in wheat, while barley for producing beer will take into account the screening percentage, malt extract, and diastatic power, which in turn is related with nitrogen content.

This article aims to summarize key elements of grain structure, grain growth, and synthesis of major grains components in field crops in order to highlight the main attributes which modify grain quality.

Grain Structure

Harvested cereal and oilseed organs may comprise true seeds (soybeans, rapeseed) or fruits (seeds and maternal-accompanying structures, like sunflower achenes or wheat, barley, rice, maize, and sorghum caryopses). Seeds develop from fertilized ovules and consist of three genetically different tissues: (a) the embryo



Grain Quality in Oil and Cereal Crops. Figure 1

Schematic postharvest processing and storage of wheat and barley production and main quality attributes in each step

developed from a zygote (diploid, representing the next generation), (b) the endosperm (usually triploid), and (c) the seed coat formed out by integuments, representing the maternal tissues of the ovule [3]. The proportion of these three components differs in mature seeds of cereals and oilseeds; endosperm is preponderant in cereals while the embryo prevails in oilseeds. With a few exceptions, the development of the endosperm always precedes that of the embryo; and the seed coat development precedes both. These genetically different parts interact closely during development and germination, and recent studies demonstrate the complexity of the connections and regulations among the different seed tissues [4, 5]. After fertilization and seed setting, grains are the primary sink in the plant. Grain filling requires important amounts of photoassimilates supplied by the mother plant through actual photosynthesis and/or the remobilization of stored carbohydrates from vegetative structures. No vascular connection exists between the mother plant and the developing embryo [5, 6] so grain growth is therefore sustained by water and solute movement through cell membranes regulated by both mother plant and seed.

Seed-attached structures include coats (testa and tegmen) and other diverse maternal-originated

structures, like the lemma and palea in cereals, pods in soybeans, siliques in rapeseed, and hull (ovary wall attached to the floral receptacle) in sunflower. These structures can greatly influence grain quality appreciation. The seed coat color in different types of beans (*Phaseolus*) impacts consumers differently according to the region, causing rejection of some genotypes albeit their good nutritional properties. Sorghum caryopsis with or without tannins are another example of the importance of grain coats affecting seed quality. Some seed coats can provide nutrients, like the B-group vitamins and micronutrients in cereal brans. In addition, they contribute to other important biological and technological functions, protecting the seed from mechanical damage in postharvest, or by affecting the industrial grain processing (wheat grinding, barley malting, rice parboiling). Seed coats can also impact seed dormancy and germination processes [7]. During recent years, seed-attached structures have received special attention as influencing the potential grain size and volume [8–11].

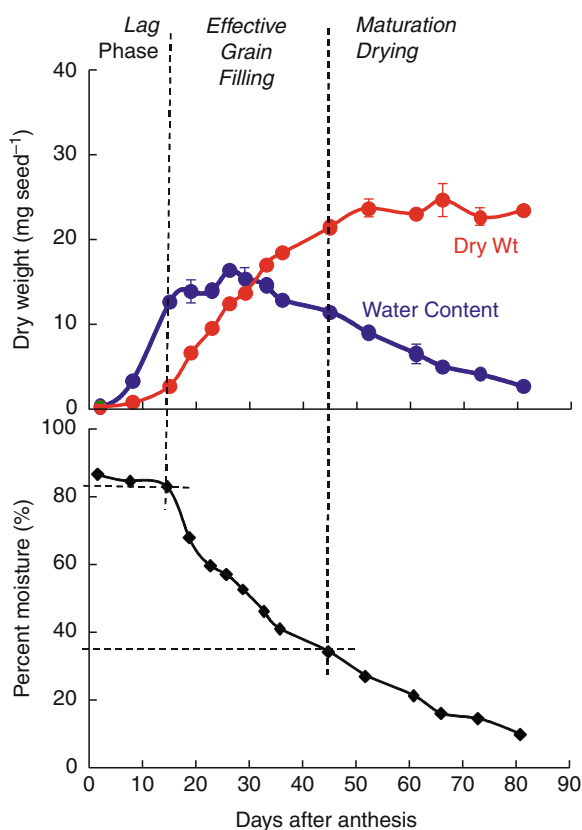
Seeds store carbohydrates (starch, oil) and proteins (soluble and insoluble). The places where these reserves are accumulated vary widely between cereals and oilseeds. In cereals, the tissue that specializes in storing

starch and protein is the endosperm. In contrast, oil seeds do not have a specialized storage tissue; oil and protein accumulate in embryo and cotyledon cells. The well-developed starchy endosperm of cereals, with an outer aleurone layer, can comprise as much as 80% of the dry weight of the mature seed. The mature endosperm consists of dead cells packed with starch granules embedded in a protein matrix. The embryo is relatively small, accounting for only about 1–2% of the seed dry weight in wheat, and is usually located on one side of the seed near the point of attachment of the seed to the mother plant [3, 6]. The non-endospermic true seed of oilseeds consists of a large embryo with two cotyledons and the embryo axis. The majority of the reserve materials are stored in the cotyledons, which make up as much as 70% (sunflower) to 90% (soybean, rapeseed) of the total seed dry weight [6].

Grain structure is important since it determines grain and industrial processing quality. Cereal endosperm structure is defined by the number, shape, and size of the starch granules, together with the quantity and type of proteins in the protein matrix. Endosperm structure is used to classify wheat according to its hardness (soft, hard), thus affecting its industrial processing quality (milling capacity and flour yield). In addition, endosperm structure is used to separate dent and flint maize according to the quantity and partitioning of the floury and horny endosperm (greater proportion of horny endosperm in flint maize). Other endosperm structure characteristics that affect grain quality are vitreousness and color, both important for maize, rice, and bread and pasta wheat. Grain structure is also important for defining oilseed quality. In sunflower, the proportion of hull and embryo is an important attribute that defines oil yield, since the hull does not store oil and therefore reduces the oil concentration in the embryo. In the past 30 years, genetic improvement has reduced the hull proportion of sunflower oilseed, increasing the oil percentage on the whole seed [12]. However, thin hulls are usually harder to remove during industrial processing, so other improvement strategies are needed to increase the percentage of sunflower oil in the future. Grain structure has, therefore, a strong impact on the commercial and industrial quality of the grain, and for this reason its attributes are present in grain marketing regulations worldwide.

Grain Growth and Source–Sink Balance

Seed Biomass During grain filling, the pollinated flower undergoes cell division and differentiation and forms a mature grain (development), which increases in size and weight (growth), reaching mature grain dry weights of 30–50 mg (wheat-barley), 250–400 mg (maize), 20–25 mg (rice), 30–50 mg (sunflower), 150–400 mg (soybeans), and 2–5 mg (rapeseed). Growth and development dynamics can be described by analyzing the rate and time period of grain growth (Fig. 2). The latter are useful tools to explain changes in the final grain weight due to genotypic and environmental factors. Species differ in their biomass per seed, and ample intra-specific differences are also observed [6]. Commercial genotypes used by farmers in maize, wheat, and soybean show differences in seed size, and this variability is even larger when exotic material is considered.



Grain Quality in Oil and Cereal Crops. Figure 2 Dynamics of individual seed dry weight (Dry Wt), water content per seed, and seed moisture of wheat seed

Seed biomass accumulation is commonly partitioned into three phases: the lag phase, the effective seed-filling period, and the maturation drying phase (Fig. 2). The lag phase is a period of active cell division. It is characterized by a rapid increase in water content with almost no dry matter accumulation. Following the lag phase, cells within the seed enter a differentiation and maturation phase, and a period of rapid dry matter accumulation resulting from the deposition of seed reserves. This phase is generally referred to as the effective seed-filling period. As in the lag phase, water content continues to increase rapidly and eventually establishes the maximum volume of the seed. Species vary considerably as to when maximum seed water content is achieved during seed filling [13]. In maize kernels, maximum water content occurs near mid seed filling [14], while in soybean seeds maximum water content is achieved at a later stage, when 70–80% of the final seed size has been achieved [15] and conversely, sunflower reaches it earlier with only 30% of final grain dry weight [10]. During the third phase of development, seeds lose water content, reach “physiological maturity” (maximum dry matter accumulation), and enter a quiescent state [3]. Seed water concentration declines throughout the three stages of seed development (Fig. 2). This decline is most obvious after seeds reach physiological maturity, but it also occurs during rapid seed filling as water is displaced by reserves [14–16].

The progress of dry matter accumulation in developing seeds and the concurrent loss of water are closely related phenomena. Studies with maize, wheat, soybean, and sunflower [17–20] have shown that final seed size is achieved at, or near, a minimum water concentration. Also, results from several studies have shown that seed water concentration accurately predicts the percent of maximum seed size achieved at any moment during seed filling in wheat, soybean, maize, and sunflower [17–20]. Such results support the notion that the duration of seed filling is determined by the interaction between reserve deposition and declining cellular water content, where deposition of reserves such as starch, protein, or lipids replace water until a critical minimum water concentration is reached [6, 20, 21]. Species differ in the seed water concentration when they achieve maximum seed biomass [13]. For example, soybean seeds reach

maturity at ~62%, maize seeds at ~36%, and wheat seeds at ~37% moisture. Although minor compared to differences across species, it has been shown that when an ample set of cultivars within a species is analyzed, variability for this trait can also be observed [22].

The rate of seed growth during the effective seed filling is highly dependent upon the number of sites for reserve deposition. The usual estimate of seed sink capacity is the number of differentiated cells during the lag phase. In maize, wheat, and other cereals, the number of endosperm cells is highly related to the rate of seed growth during rapid seed filling. In legumes such as soybean or pea, the number of cotyledon cells is highly related to the rate of seed growth. Thus, rate and duration of grain filling are important to define the final grain weight, an important attribute of grain quality.

Source–Sink Balance In higher plants, nutrients from assimilation sites (sources) are delivered to sites of nutrient utilization (sinks) through an interconnected network of sieve elements. Partitioning of phloem-delivered nutrients between competing sinks is governed by their relative ability to unload major osmotic species from the importing phloem sieve elements [23]. This process depends upon a set of intercellular (post-sieve element) transport events which are integrated with growth or storage functions of the recipient sink tissues [24].

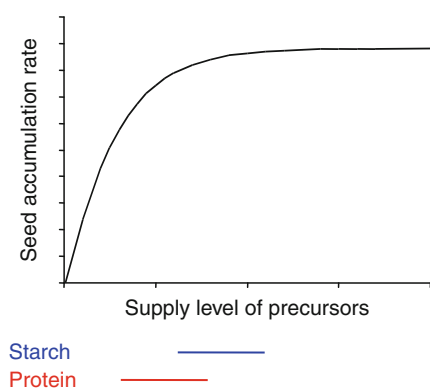
Species differ on the seed size at maturity [25], and this interspecific variability is more related to the amount of assimilates available per seed during the early lag phase than during the effective seed-filling period [21]. At flowering, plants adjust the number of seeds and the potential seed size to the growth environment [21], and species differ in how they distribute available assimilates into more seeds or more potential seed size at around the period when seed number is being determined [26]. Seed size is mainly determined by the genotype, although the environment can affect the final size as well. Water availability and temperature are two environmental conditions that can create important changes in the size of the seeds at maturity.

The amount of assimilates available per seed is usually referred to as the source–sink balance, and is used to describe the relation between the total amount of available assimilates and the sink number. This ratio is used

to simplify the idea of assimilate availability per sink, and the way the source–sink ratio has been estimated can vary widely. Different researchers have used plant growth per seed, green leaf area per seed, sucrose availability per seed, and alternative approaches including plant growth per day per unit of sink growth per day. The source–sink balance that the seeds experience during their growth is adjusted at around flowering, when plants are setting the number of seeds.

Because plants grow in a nonuniform environmental condition, the source–sink balance during the period when seeds are accumulating biomass can change. An example can be a defoliation caused by an insect eating leaves attacking the crop at mid grain filling (which would reduce the source–sink balance of the crop) or a drought stress reducing plant growth (also reducing the source–sink balance). The source–sink balance becomes relevant because not all seed components vary to the same degree when assimilate availability per seed is altered, so the seed composition and quality may change [27–29].

Jenner and coworkers developed a theoretical model to understand how changes in the amount of assimilates available per seed can affect seed composition [28]. Their model is based on the idea that each one of the seed components can be more or less affected by changes in the level of precursors available for the growing seed because not all components are receiving from the mother plant the same level of precursors needed for their deposition within the seed. An example is illustrated in Fig. 3, where changes in the level of precursors



Grain Quality in Oil and Cereal Crops. Figure 3 Level of precursors available per seed to synthesis of different components of seeds (Adapted from [28])

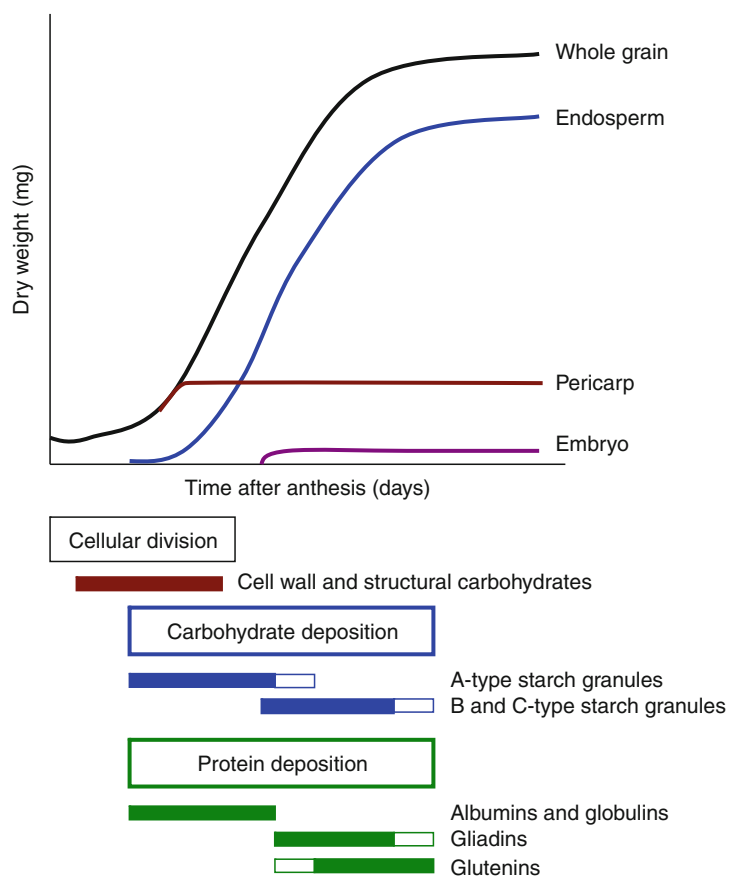
will most surely not affect the starch content of the seed but changes in the level of precursors needed for protein synthesis will affect the protein content and the final protein concentration. This model helps explain why changes in plant growth (that affect the precursor levels) will most surely not affect starch content but will surely affect the protein content of the seed. Recent studies conducted on soybean [30] and maize compositions [27, 31] agree with this model.

Synthesis of Major Components

Grain growth involves growth processes of various structures (seed coats, embryo, endosperm) and accumulation of different substances (starch, oil, protein). Grain components are not synthesized simultaneously nor do they occur at the same rate; thus, physical and chemical grain composition varies during grain filling. This is an important aspect when dealing with industrial and nutritional grain quality.

As an example, the different parts of a wheat grain and the main components synthesized during grain filling are shown in Fig. 4. Starch is the main component in wheat, comprising 70–80% of the final grain weight. Different types of starch granules of varying numbers, shapes, and sizes are synthesized into the endosperm cells. The A-type granules are bigger and quantitatively more important than the succeeding B- and C-type granules. Starch is composed of amylose and amylopectin at a 3:1 ratio, except in waxy genotypes where amylopectin is more abundant.

In wheat bread, proteins comprise 5–20% of the final grain dry weight and include albumins and globulins (30–40%); gliadins and glutenins (60–70%). Albumins and globulins have enzymatic and metabolic functions and are located in the embryo and aleurone layer; gliadins and glutenins form the gluten and are reserve proteins, confined in the endosperm [32]. During grain filling, metabolic proteins are synthesized first and predominate until 10–15 days after anthesis. Reserve proteins accumulate during the effective filling phase; gliadins are the first to be detected (10–15 days after anthesis) while glutenins are deposited later (15–20 days after anthesis). Gliadins give viscosity to the mass while the glutenins confer elasticity, and both result in viscoelastic gluten appropriate for a good loaf volume. Since the gliadins:glutenins ratio changes



Grain Quality in Oil and Cereal Crops. Figure 4

Grain filling dynamics of a cereal (wheat) and synthesis of major grain components (Adapted from [29])

during grain filling, crop exposure to stressful conditions (i.e., high temperature, water stress) during this phase will modify the total protein mass and the gliadins:glutenins ratio, affecting baking quality.

Grains from other species have different grain component synthesis patterns. For example, in maize there is only one type of starch granule and although starch is synthesized through the entire effective filling phase amylose accumulation occurs after amylopectin. Reserve proteins (5–14% of the final grain weight) are accumulated during the entire effective filling period as well, forming protein bodies in the endosperm cells. Oil accumulation (3–15% of the final grain weight) takes place at the end of the filling phase and most of the oil is found in the embryo. In oilseeds, oil and protein are the main reserve substances; carbohydrates are scarcely accumulated (<20%). Sunflower seeds contain 40–55% oil and 10–20% protein, while soybean seeds

contain greater protein percentages (35–50%) and lower oil percentages (10–25%). Both protein and oil are deposited in the embryo cells during the linear grain filling phase. Oil deposits form oleosomes or lipid bodies of spherical shapes, while reserve proteins form dense and irregular protein bodies [33, 34].

During grain filling oilseeds, protein synthesis usually occurs after oil synthesis. Oil is formed by triglycerides, which are composed of one glycerol molecule combined with three fatty acids. Fatty acids differ in the number of carbon atoms (typically between 14 and 22 in vegetable oils) and the number of double bonds between carbon atoms. The different proportions of fatty acids modify the physicochemical and industrial properties of oils. Oils with high saturated fatty acid percentages (without double bonds; like palmitic and stearic acids) are semisolid at room temperature, with a high melting point and a higher resistance to

oxidation (fat degradation due to oxygen presence). However, consumption of these oils (especially palmitic acid) increases cholesterol levels in the bloodstream. On the contrary, oils with high proportions of monounsaturated fatty acids (like oleic acid) and polyunsaturated fatty acids (like linoleic and linolenic acids with two and three double bonds, respectively) are liquid at room temperature, have a lower melting point, and a higher susceptibility to oxidation as the number of double bonds increase; these oils are healthier than saturated fatty acids. During grain filling, the proportion of fatty acids varies according to the species and crop varieties. In traditional sunflower genotypes, the oleic:linoleic ratio decreases during grain filling while total oil accumulation increases. In contrast, in “high oleic” sunflowers, the oleic proportion is high and constant during grain filling due to the low activity of the enzyme responsible for the linoleic synthesis deriving from oleic acid [35, 36]. Oil final composition varies greatly among oilseed species, and genetic improvement has achieved a wide variety of fatty acid compositions within the same species resulting from physical and chemical mutagenesis that affect specific enzyme functions responsible for the presence of double bonds in fatty acids [37, 38]. These enzymes are also affected by the environment, producing changes in grain oil content and composition.

Main Factors Affecting Grain Quality

Genotypic Effects

Some few grain attributes are mainly driven by the genotype, and the environment has relatively low influence. For example, the color of the wheat grain (white, yellow) is strongly determined by the ability of genotype to accumulate lutein, and the character “high oleic” in oilseed genotypes is associated with genetic mutations defective for the enzyme that desaturates oleic acid. Genotypes within the same species can present large differences in grain composition. For example, in commercial genotypes of soybean, the difference in protein percentage can easily vary from 33% to 42% [39]. This variability can be observed for any seed component in any species [40–42], as ample natural variation is common (Table 1).

Earlier studies working on understanding the genetic basis of genotypic differences in seed

Grain Quality in Oil and Cereal Crops. Table 1 Grain composition ranges reported in different species

Species	Grain composition range	Source
Soybean	33–42% protein	[35]
Sunflower	20–30% oil (confectionary type)	[33, 38]
	40–55% oil (oil type)	
Wheat bread	5–20% protein	[40]
Corn	5–14% protein	[31]
	8–12% protein (pop corn type)	[41]
	5–10% oil (high oil content)	[42]

composition were based on the use of mutants, and these usually yielded qualitative differences in seed composition. For example, a commercial genotype of maize usually contains ~40% amylopectin and ~60% amylose, and a waxy mutant contains 100% amylopectin and no amylose. At present time, a large number of mutants have been discovered and used within any species as specialty quality genotypes.

The modification of the fatty acid profile of oil seeds has been one of the main tasks faced by oilseed breeders over the past 40 years. Success in this field has been of paramount importance for the worldwide expansion of some oilseed crops. The elimination of erucic acid (a harmful fatty acid) from rapeseed oil was the first step toward the development of canola (zero-erucic, low-glucosinolate rapeseed) as one of the major sources of vegetable oil in the world. Other landmarks in oilseed breeding for seed oil quality have been the development of high oleic, low linolenic acid canola, low linolenic acid linseed and soybean, high oleic acid sunflower, high saturated sunflower, and sunflower lines with modified tocopherol (antioxidant compounds) composition. Most of these traits defining seed oil quality have been found to be governed by a reduced number of genes (one to three major genes, with several alleles for each locus in most cases), and this fact implies that the practical management of single quality traits in breeding programs is relatively easy if compared with polygenic traits (as grain yield, grain weight or protein and oil content). Additionally, the fatty acid composition of the seed oil is determined by the genotype of the developing embryo (not the whole plant), so mutagenesis and selection can be

carried out at a single-seed level, using the half-seed technique.

In wheat, improving yield potential without negatively affecting grain quality is difficult, mainly because increases in grain yield are generally accompanied by a decrease in grain protein content, which is strongly associated with bread-making quality. Wheat breeders give grain quality the same level of importance as yield potential and disease resistance. In contrast to the low heritability of protein content, grain hardness and yellow pigment are highly heritable and can be readily improved through conventional breeding. Plant breeders select at least one parent with the desired quality when designing their crossing strategies, particularly as end-use requirements frequently determine the fate of potential new cultivars, but the stage in the breeding process at which quality determination takes place will influence which tests (micro or macro tests) are applied, according to the sample size available.

At present, natural variation for seed composition is being studied identifying quantitative trait loci (QTLs) for different seed components (oil, protein), as any seed component is a quantitative trait governed by many genes and each one with an individual small effect. The study by Blanco and coworkers can be mentioned as an example, where the authors studied seed protein concentration in wheat where three major QTLs were detected [43]. This methodology is currently becoming very popular and has yielded molecular markers associated with seed component traits that can help understand the genetic bases of the trait and be used by breeders and the industry.

Environmental Effects

As mentioned earlier, the majority of quality traits are greatly modified by the environment and by genotype–environment interactions. Grain weight and protein concentration are found within this group of traits. Environment variables like high temperature, water and nitrogen (N) availabilities have been the most studied modifying grain quality.

The response in grain composition to a particular stress depends mainly on the stress characteristics (i.e., intensity, duration of the stressful period, opportunity of occurrence, and the interaction that this stress may have with other stresses to which the crop is exposed

to). The relevance of the intensity and duration of the stress on the magnitude of the change in grain composition is self-explained (the more severe and the longer a stress is, the greater change in composition produced, though not necessarily this implies that the relationship is linear). The timing of occurrence is also critical, as shown in Fig. 4 not all stages are equally critical for the final determination of grain quality: if the stress coincides with critical stages for synthesis and deposition of the components, the changes will result far stronger than that of stresses occurring in less-critical stages. Therefore, the grain composition responses to stressful factors may range from virtual insensitivity (if punctual synthesis reductions are compensated by recovering when no stress occurs) to different ranges of quality reductions to even crop failure to produce a certain quality level.

Seed growth and development are responsive to temperature, but their responses vary with the temperature range considered [44]. As a general rule, the rate of seed development increases as temperature increases, reducing the duration of seed-filling period. At lower temperatures, seed growth rates decrease linearly as temperatures fall below 15°C in wheat, soybean, rice, sunflower, and maize. Seed growth rates increase when temperatures rise from 20°C to 30°C; however, this increase does not offset the linear decrease in seed-filling duration, resulting in lower grain weights [44]. In most cases, moderately high temperatures (20–30°C) prevail during grain filling, although short periods of very high temperatures (>30–32°C) may occur reducing seed growth rate and causing the early end of grain filling period. In addition, the earlier the heat stress, the greater the impact on grain weight [45, 46]. Brief periods of high temperatures can cause reductions in grain weight, but these effects can be overlooked if only the average temperature during post-flowering period is considered. Thus, moderately high temperatures (20–30°C) during the post-flowering period reduced grain weight mainly through shortening the grain filling period, while very high temperatures (>30–32°C) even for a few days can reduce grain weight by reducing grain filling rate and the early cessation of grain growth period. Both aspects of post-flowering temperature should be considered especially because climatic change could bring about high-temperature scenarios

in the next decades, together with an increase in heat-stress events [47, 48].

Grain quality and composition are also affected by temperature. Several experiments suggest that the temperature effects on seed composition are related to dry matter metabolism and accumulation. The timing, intensity, and duration of occurrence of heat stress may alter final grain quality according to the grain component synthesis process involved (carbohydrates, proteins, oils). Interestingly, there are some reports on the possibility of recovery post-stress [49, 50]. In wheat and barley, protein percentage increases with increasing temperatures (15–30°C) because the negative impact of high temperatures on starch synthesis is greater than the impact on protein synthesis, thus decreasing the starch proportion in the grains [28]. High temperatures also affect protein quality, generally increasing gliadin:glutenin ratio, which causes weak dough with a low bread-making quality. The temperature impact on wheat grain quality will therefore depend on the balance between the positive (higher protein) and negative (greater gliadin:glutenin ratio) effects. Temperature also affects oil fatty acid composition in oilseeds [51]. The higher the temperatures during grain filling, the higher the fatty acid saturation (i.e., greater proportions of oleic acids and lower proportions of linoleic and linolenic acids) due to the reduced activity of unsaturation enzymes in grains [52]. Temperatures registered during the night in early grain filling phases have shown to have the best predictive values for modeling the final oil composition in sunflower [53]. Progress in modeling the quality of other grains is underway [54].

In field crops, high-temperature occurrences are commonly associated to water stress, increasing the negative temperature effects. Drought stress produces a shortage of assimilates and often reduced N availability, which cause a reduction in grain growth. In general, a drought episode occurring after flowering has a similar effect as an increase in temperature – the quantity (mg grain^{-1}) of protein per grain remains stable, while starch accumulation in grain is significantly reduced, resulting in smaller grains with a greater protein percentage [55]. In oilseeds, post-flowering droughts decrease grain oil percentages and increase protein percentages [56, 57] indicating that carbon metabolism is affected to a greater extent

than N metabolism. Water stress has a smaller impact on fatty acid composition; in general, droughts do not modify the saturation degree in oils except under severe stress conditions which produce an early grain-filling cessation [58].

N availability also affects final cereal and oilseed grain composition. In general, when soil N availability is low, cereal crop yields respond positively to N fertilization. A dilution effect occurs when N taken up by the crop is partitioned in a greater number of grains, which reduces grain protein percentage. If N availability is further increased, both crop yield and grain protein percentage are increased. In addition, the stage of development when N is added is important in defining wheat grain quality. N applications around flowering increase nitrogen availability per grain, increasing protein percentage. It is reported that increases in N availability result in increases of gliadin:glutenin ratio, which in turn produce a weakening of the dough [59]. In oilseeds, a greater soil N availability increases crop yield and grain protein percentage. Consequently, oil percentages in grain decrease due to the negative relationship between oil and protein (expressed as a percentage of the grain weight). Nitrogen application effects on the grain fatty acid composition are smaller and more variable compared to temperature and water stress effects [60, 61]. A greater knowledge on the physiological processes that regulate the responses to these environmental factors is essential to decide the management of the crop to produce grain for a specific end use.

Management Strategies

Although both grain yield and quality are determined throughout the growing season, important decisions that will strongly affect them should be taken before planting [62]. The farmer's choice of genotype and the amount of nitrogen available are central for successfully combining the genotype potential for yield and quality with the environmental availability of resources. As stated earlier, final grain quality is the result of the interaction between the genotype, the natural environment, and the crop management practices [63]. In extensive production systems, it is not possible to provide each stage of the crop cycle with the optimal combination of environmental factors to reach

the highest possible yield and quality, therefore, a trade-off is to make preplanting decisions to ensure that critical crop stages for the definition of yield and quality are given a preferential environment [62]. Nevertheless, knowledge of the effects of environment and $G \times E$ interaction is still rather imprecise, so management strategies with the objective of increasing yields, while obtaining high quality, are difficult to design.

There are a number of grain quality attributes that are strongly governed by the genotype and therefore choosing the proper genotype in relation to the final end use of the grain is critical. In several countries, for trading purposes wheat is classified into distinct categories of endosperm hardness (soft, semihard, and hard). Grain hardness is determined by the packing of grain components in the endosperm cells [40] and according to this attribute, the end product can vary from pasta (hard endosperm), biscuits (soft endosperm), to bread (hard endosperm). Usually, this classification can be more detailed and complex [64]. In the case of sunflower, oil fatty acid composition is genetically controlled [65], and the oil composition has been modified mostly by altering the function of major genes through mutagenesis [38].

Addition of nitrogen fertilizer is one of the most frequent management practices for altering grain quality (and of course grain yield). It is difficult a priori to know the effect of adding nitrogen to grain quality as many other factors are intervening and modify the final expected result. In the case of wheat crops, the initial amount of nitrogen in the soil, the specific moment of fertilization, the amount of available water, and rain pattern during the growth cycle as well as plant density at sowing and genotype nitrogen use efficiency are the main factors that interact and may modify the final response in grain quality. In general, it is accepted that regardless of the species, the increase in grain yield leads to a decrease in the protein to starch or oil ratio. This negative relationship between yield and grain N concentration reflects the fact that carbon assimilation and accumulation during the grain filling period is sink-limited [66] while nitrogen accumulation in grains is usually source-limited [67], as a result of dilution effects. The final protein concentration will thus depend on the balance between the source capacity to provide nitrogen and the strength of the sink for accumulating carbohydrates [68].

Future Directions

The compositional requirements for a particular grain vary from one product to the other depending on its end use. In addition, grain quality is a dynamic concept as it changes constantly as new uses can be developed for particular grains. The three major pillars of grain composition are: the genotypes, the environments during grain growth, and their interaction.

On the genetic pillar, the knowledge gained in the recent past has been extraordinary. Based on the molecular tools developed, a number of genes and QTLs involved in the determination of particular grain components (in turn determining grain quality attributes) have been identified and mapped in several crops, and it seems easy to predict that in the near future almost any breeding program in the world will be able to manipulate these genetic factors with certainty.

Regarding the environment during grain filling, important and useful findings have been reported in relation to high temperatures, and in lesser extent in water stress, and nitrogen availability. Few studies have attempted to examine the interactions between these environmental factors on grain quality attributes. It has been recently reported that high-temperature stress effects may be mitigated under high nitrogen availability for wheat and barley [69–71].

Undoubtedly, the challenge for breeders and agronomist is dealing with $G \times E$ interactions [72]. Therefore, there is a need for increasing current knowledge on the physiology of quality traits in order to obtain both high yield and high quality through breeding and management strategies. This will also help predict grain composition through a series of genotypes and environments.

Using agronomic simulation models properly calibrated and validated for the target population of environments can be a tool for understanding and predict final grain composition. The incorporation of grain quality modules into crop simulation models is increasing (European Journal of Agronomy 25, 2006). Grain protein content was the first trait incorporated into modeling as well as grain size (grain weight) which is a quality criteria especially valued by millers in the case of cereals but also for oil extraction in oil crops. Recently, more detailed concepts have been incorporated such as

the type of protein [73] and oil quality [74]. It is expected that incorporating genetic data into simulation routines will be done in the near future.

Bibliography

Primary Literature

1. Slafer GA, Satorre EH (1999) Wheat production systems of the Pampas. In: Satorre EH, Slafer GA (eds) *Wheat: ecology and physiology of yield determination*. Food Product Press, New York, pp 333–343
2. Wrigley CW (1994) Developing better strategies to improve grain quality for wheat. *Aust J Agric Res* 45:1–7
3. Boesewinkel FD, Bouman F (1995) The seed: structure and function. In: Kigel J, Galili G (eds) *Seed development and germination*. Marcel Dekker, New York, pp 1–24
4. Berger F, Grini PE, Schnittger A (2006) Endosperm: an integrator of seed growth and development. *Curr Opin Plant Biol* 9:664–670
5. Meyer CJ, Steudle E, Peterson CA (2007) Patterns and kinetics of water uptake by soybean seeds. *J Exp Bot* 58:717–732
6. Egli DB (1998) *Seed biology and the yield of grain crops*. CAB International, New York, 178 p
7. Baskin JM, Baskin CC (2004) A classification system for seed dormancy. *Seed Sci Res* 14:1–16
8. Millet E, Pinthus MJ (1984) The association between grain volume and grain weight in wheat. *J Cereal Sci* 2:31–35
9. Calderini DF, Abledo LG, Slafer GA (2000) Physiological maturity in wheat based on kernel water and dry matter. *Agron J* 92:895–901
10. Rondanini DP, Mantese AI, Savin R, Hall AJ (2009) Water content dynamics of achene, pericarp and embryo in sunflower: associations with achene potential size and dry-down. *Eur J Agron* 30:53–62
11. Lizana XC, Riegel R, Gomez LD, Herrera J, Isla A, McQueen-Mason SJ, Calderini DF (2010) Expansins expression is associated with grain size dynamics in wheat (*Triticum aestivum* L.). *J Exp Bot* 61:1147–1157
12. Putt ED (1997) Early history of sunflower. In: Schneiter AA (ed) *Sunflower technology and production*. American Society of Agronomy, Madison, pp 1–19
13. Egli DB, TeKrony DM (1997) Species differences in seed water status during seed maturation and germination. *Seed Sci Res* 21:289–294
14. Westgate ME, Boyer JS (1986) Water status and the developing grain of maize. *Agron J* 78:714–719
15. Egli DB (1990) Seed water relations and the regulation of the duration of seed growth in soybean. *J Exp Bot* 41:243–248
16. Borrás L, Westgate ME, Otegui ME (2003) Control of kernel weight and kernel water relations by post-flowering source-sink ratio in maize. *Ann Bot* 91:857–867
17. Borrás L, Westgate ME (2006) Predicting maize kernel sink capacity early in development. *Field Crop Res* 95: 223–233
18. Swank JC, Egli DB, Pfeiffer TW (1987) Seed growth characteristics of soybean genotypes differing in duration of seed fill. *Crop Sci* 27:85–89
19. Rondanini DP, Savin R, Hall AJ (2007) Estimation of physiological maturity in sunflower as a function of fruit water concentration. *Eur J Agron* 26:295–309
20. Schnyder H, Baum U (1992) Growth of the grain of wheat (*Triticum aestivum* L.): the relationship between water content and dry matter accumulation. *Eur J Agron* 1:51–57
21. Gambin BL, Borrás L (2010) Resource distribution and the trade-off between seed number and weight: a comparison across crop species. *Ann Appl Biol* 156:91–102
22. Borrás L, Zinselmeier C, Senior ML, Westgate ME, Muszynski MG (2009) Characterization of grain filling patterns in diverse maize germplasm. *Crop Sci* 49:999–1009
23. Patrick JW (1997) Phloem unloading: sieve element unloading and post-sieve element transport. *Annu Rev Plant Biol* 48:191–222
24. Patrick JW, Offler CE (2001) Compartmentation of transport and transfer events in developing seeds. *J Exp Bot* 52:551–564
25. Egli DB (1981) Species differences in seed growth characteristics. *Field Crop Res* 4:1–12
26. Sadras VO (2007) Evolutionary aspects of the trade-off between seed size and number in crops. *Field Crop Res* 100:125–138
27. Borrás L, Curá JA, Otegui ME (2002) Maize kernel composition and post-flowering source-sink ratio. *Crop Sci* 42:781–790
28. Jenner CF, Ugalde TD, Aspinall D (1991) The physiology of starch and protein deposition in the endosperm of wheat. *Aust J Plant Physiol* 18:211–226
29. Savin R, Molina-Cano JL (2002) Changes in malting quality and its determinants in response to abiotic stress. In: Slafer GA, Molina-Cano JL, Savin R, Araus JL, Romagosa I (eds) *Barley science: recent advances from molecular biology to agronomy of yield and quality*. Food Product Press, New York, pp 523–544
30. Rotundo JL, Borrás L, Westgate ME, Orf JH (2009) Relationship between assimilates supply per seed and soybean seed composition. *Field Crop Res* 112:90–96
31. Seebauer JR, Singletary GW, Krumpelman PM, Ruffo ML, Below FE (2010) Relationship of source and sink in determining kernel composition of maize. *J Exp Bot* 61:511–519
32. Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: structures and biosynthesis. *Plant Cell* 7:945–956
33. Mantese AI, Medan D, Hall AJ (2006) Achene structure, development and lipid accumulation in sunflower cultivars differing in oil content at maturity. *Ann Bot* 97:999–1010
34. Tanaka W, Mantese AI, Maddonni GA (2009) Pollen source effects on growth of kernel structures and embryo chemical compounds in maize. *Ann Bot* 104:325–334
35. Garcés R, Mancha M (1989) Oleate desaturation in seeds of two genotypes of sunflower. *Phytochemistry* 28:2593–2595
36. Ohlrogge J (1997) Regulation of fatty acid synthesis. *Annu Rev Plant Physiol Plant Mol Biol* 48:109–136
37. Harwood JL (1996) Recent advances in the biosynthesis of plant fatty acids. *Biochim Biophys Acta* 1301:7–56

38. Velasco L, Perez-Vich B, Fernández-Martínez JM (2004) Grain quality in oil crops. In: Benech-Arnold RL, Sánchez RA (eds) Handbook of seed physiology: applications to agriculture. Food Products Press/The Haworth Press, New York, pp 389–405
39. Rotundo JL, Westgate ME (2009) Meta-analysis of environmental effects on soybean seed composition. *Field Crop Res* 110:147–156
40. Peña RJ, Trethowan R, Pfeiffer WH, van Ginkel M (2002) Quality (end-use) improvement in wheat compositional, genetic, and environmental factors. *J Crop Prod* 5:1–37
41. Park D, Allen KGD, Stermitz FR, Maga JA (2000) Chemical composition and physical characteristics of unpopped popcorn hybrids. *J Food Comp Anal* 13:921–934
42. Thomison PR, Geyer AB (1999) Evaluation of TC-Blend7 used in high oil maize production. *Plant Var Seeds* 12:99–112
43. Blanco A, Simeone R, Gadaleta A (2006) Detection of QTLs for grain protein content in durum wheat. *Theor Appl Genet* 112:1195–1204
44. Wardlaw IF, Wrigley C (1994) Heat tolerance in temperate cereals: an overview. *Aust J Plant Physiol* 21:695–703
45. Savin R, Nicolas M (1999) Effects of timing of heat stress and drought on growth and quality of barley grains. *Aust J Agric Res* 50:357–364
46. Stone PJ, Nicolas ME (1996) Effect of timing of heat stress during grain filling on two wheat varieties differing in heat tolerance. II. Fractional protein accumulation. *Aust J Plant Physiol* 23:739–749
47. Easterling D, Horton B, Jones P, Peterson T, Karl T, Parker D, Salinger M, Razuvayev V, Plummer N, Jamason P, Folland C (1997) Maximum and minimum temperature trends for the globe. *Science* 277:364–367
48. Meehl GA, Tebaldi C (2004) More intense, more frequent, and longer lasting heat waves in the 21st century. *Science* 305:994–997
49. Hawker JS, Jenner CF (1993) High temperature affects the activity of enzymes in the committed pathway of starch synthesis in developing wheat endosperm. *Aust J Plant Physiol* 20:197–209
50. Stone PJ, Gras PW, Nicolas ME (1997) The influence of recovery temperature on the effects of a brief heat shock on wheat. III. Grain protein composition and dough properties. *J Cereal Sci* 25:129–141
51. Canvin D (1965) The effect of temperature on the oil content and fatty acid composition of the oils from several oil seed crops. *Can J Exp Bot* 43:63–69
52. Garcés R, Mancha M (1991) In vitro oleate desaturase in developing sunflower seeds. *Phytochem* 30:2127–2130
53. Izquierdo N, Aguirrezábal LAN, Andrade F, Cantarero M (2006) Modeling the response of fatty acid composition to temperature in a traditional sunflower hybrid. *Agron J* 98:451–461
54. Martre P (2006) Modelling quality traits and their genetic variability for wheat. *Eur J Agron* 25:75–78
55. Triboi E, Martre P, Triboi-Blondel AM (2003) Environmentally-induced changes in protein composition in developing grains of wheat are related to changes in total protein content. *J Exp Bot* 54:1731–1742
56. Champolivier L, Merrien A (1996) Evolution de la teneur en huile et de sa composition en acides gras chez deux variétés de tournesol (oléique ou non) sous l'effet de températures différentes pendant la maturation des graines. *Oleagineux Corps Gras Lipides* 3:140–145
57. Rotundo JL, Westgate ME (2010) Rate and duration of seed component accumulation in water-stressed soybean. *Crop Sci* 50:676–684
58. Flagella Z, Rotunno T, Tarantino E, Di Caterina R, De Caro A (2002) Changes in seed yield and oil fatty acid composition of high oleic sunflower (*Helianthus annuus* L.) hybrids in relation to the sowing date and the water regime. *Eur J Agron* 17: 221–230
59. Payne PI, Holt LM, Worland AJ, Law CN (1982) Structural and genetical studies on the high-molecular-weight subunits of wheat glutenin. *Theor Appl Genet* 63:129–138
60. Steer BT, Seiler GJ (1990) Changes in fatty acid composition of sunflower (*Helianthus annuus*) seeds in response to time of nitrogen application, supply rates and defoliation. *J Sci Food Agric* 51:11–26
61. Zheljazkov VD, Vick BA, Baldwin BS, Buehring N, Astatkie T, Johnson B (2003) Oil content and saturated fatty acids in sunflower as a function of planting date, nitrogen rate, and hybrid. *Agron J* 101:1003–1011
62. Calderini DF, Dreccer MF (2002) Choosing genotype, sowing date and plant density for malting quality. In: Slafer GA, Molina-Cano JL, Savin R, Araus JL, Romagosa I (eds) Barley science. Recent advances from molecular biology to agronomy of yield and quality. Food Product Press/The Haworth Press, New York, pp 413–444
63. Gooding MJ, Davies WP (1997) Wheat production and utilization. Systems, quality and the environment. CAB International, Wallingford, 355 p
64. Wrigley CW, Bekes F (2004) Processing quality requirements for wheat and other cereal grains. In: Benech-Arnold R, Sanchez RA (eds) Handbook of seed physiology: applications to agriculture. The Haworth Press, New York, pp 389–405
65. Garcés R, Mancha M (1989) Oleate desaturation in seeds of two genotypes of sunflower. *Phytochem* 28:2593–2595
66. Borrás L, Slafer GA, Otegui ME (2004) Seed dry weight response to source-sink manipulations in wheat, maize and soybean: a quantitative reappraisal. *Field Crop Res* 86: 131–146
67. Savin R, Prystupa P, Araus JL (2006) Hordein composition as affected by post-anthesis source-sink ratio under different nitrogen availabilities. *J Cereal Sci* 44:113–116
68. Stone PJ, Savin R (1999) Grain quality and its physiological determinants. In: Satorre EH, Slafer GA (eds) Wheat: ecology and physiology of yield determination. Food Product Press/The Haworth Press, New York, pp 85–120
69. Zahedi M, Mc Donald G, Jenner CF (2004) Nitrogen supply to the grain modifies the effects of temperature on starch and

protein accumulation during grain filling in wheat. *Aust J Agric Res* 55:551–564

70. Dupont FM, Hurkman WJ, Vensel WH, Tanaka C, Kothari KM, Chung OK, Altenbach SB (2006) Protein accumulation and composition in wheat grains: effects of mineral nutrients and high temperature. *Eur J Agron* 25:96–107
71. Passarella VS, Savin R, Slafer GA (2008) Are temperature effects on weight and quality of barley grains modified by resource availability? *Aust J Agric Res* 59:510–516
72. Aguirrezábal LAN, Martre P, Pereyra-Irujo G, Izquierdo N, Allard V (2009) Management and breeding strategies for the improvement of grain and oil quality. In: Sadras VO, Calderini DF (eds) *Crop physiology: applications for genetic improvement and agronomy*. Academic/Elsevier, New York, pp 387–410
73. Martre P, Porter JR, Jamieson PD, Tribou E (2003) Modeling grain nitrogen accumulation and protein composition to understand the sink/source regulations of nitrogen remobilization for wheat. *Plant Physiol* 133:1959–1967
74. Pereyra-Irujo GA, Aguirrezábal LAN (2007) Sunflower yield and quality interactions and variability: analysis through a simple simulation model. *Agr For Meteorol* 143:252–265

Books and Reviews

- Aguirrezábal LAN, Andrade FH (1998) *Calidad de productos agrícolas. Bases ecofisiológicas, genéticas y de manejo agronómico*. Unidad Integrada INTA Balcarce, Balcarce, 315 p
- Baskin CC, Baskin JM (1998) *Seeds: ecology, biogeography, and evolution of dormancy and germination*. Academic Press/Elsevier, San Diego, 666 p
- Basra AS, Randhawa LS (2002) *Quality improvement in field crops*. Food Products Press/The Haworth Press, New York, 433 p
- Benech-Arnold RL, Sánchez RA (2004) *Handbook of seed physiology: applications to agriculture*. Food Products Press/The Haworth Press, New York, 483 p
- Bewley JD, Black M (1985) *Seeds: physiology of development and germination*, 1st edn. Plenum, New York, 125 p
- Gunstone FD, Harwood JL, Dijkstra AJ (2007) *The lipid handbook with CD-ROM*, 3rd edn. CRC Press, Boca Raton, 1472 p
- Sadras VO, Calderini DF (2009) *Crop physiology: applications for genetic improvement and agronomy*. Academic Press/Elsevier, New York, 583 p
- Schneider AA (1997) *Sunflower technology and production*. ASA, CSSA & SSSA, Madison, 834 p
- Simmonds DH (1989) *Wheat and wheat quality in Australia*. CSIRO, Melbourne, 299 p
- Slafer GA, Molina-Cano JL, Savin R, Araus JL, Romagosa I (2002) *Barley science: recent advances from molecular biology to agronomy of yield and quality*. Food Products Press/The Haworth Press, New York, 551 p
- Tribou E, Tribou-Blondel AM (2002) Productivity and grain or seed composition: a new approach to an old problem. *Eur J Agron* 16:163–186

Gravity Recovery and Climate Experiment (GRACE): Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains

VICTOR ZLOTNICKI¹, SRINIVAS BETTADPUR²,
FELIX W. LANDERER³, MICHAEL M. WATKINS³

¹Climate, Oceans and Solid Earth Science Section,
Jet Propulsion Laboratory, California Institute of
Technology, Pasadena, CA, USA

²Center for Space Research, University of Texas-Austin,
Austin, TX, USA

³Jet Propulsion Laboratory, California Institute of
Technology, Pasadena, CA, USA

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

What Is GRACE

Spherical Harmonics; Equivalent Water Thickness;
Mascons; and Spatial Resolution

Background Fields

Ambiguity; Downward Continuation; Data Accuracy

Applications in Land Hydrology

Applications in Cryospheric Studies

Applications in Ocean Studies

Applications in Solid Earth Studies

Future Directions

Acknowledgments

Bibliography

Glossary

Equivalent water thickness Since time changes in the gravity field are caused by time changes in mass distributions, equivalent water thickness (“EWT”) is the variable thickness of a thin layer of water (thin relative to both the radius of the Earth and the horizontal scale of the signals) draping the Earth that would correspond to the observed changes in gravity. The conversion from gravitational spherical harmonics to water thickness (and vice versa) is unique and well defined, regardless of what actually causes the gravitational changes. The concept is not

used when studying changes in the solid Earth, such as glacial isostatic adjustment or earthquakes.

Glacial isostatic adjustment (GIA) Also known as postglacial rebound, it is the viscoelastic response of the mantle and lithosphere to the removal of the great ice sheets that covered parts of the Earth and peaked 21,000 years ago [65]. The deglaciation was essentially complete 6,000 years ago. The lithosphere rises where the ice sheets used to be, but sinks in other locations.

Ionosphere A set of layers at altitudes between approximately 80 and 1,000 km above the Earth's surface, with electrons and electrically charged atoms. The ionosphere leads to a delay to electromagnetic radiation, which is frequency-dependent and changes with local time and solar activity. The GRACE KBR system uses two frequencies to correct for this path delay.

KBR K-band microwave ranging system measures the distance between the two GRACE satellites using two frequencies, 24 and 32 GHz.

Mascons Mass concentrations. The term was coined by Muller and Sjogren [60] to describe mass concentrations in the lunar nearside, beneath the center of the surface features termed "mare" (pl. "maria"). Today the term "mascons" refers to an alternative method to solve the GRACE gravity fields in terms of distributed spherical caps or point masses, instead of using the spherical harmonic representation.

Newton's law of gravitation It states that every point mass attracts every other point mass with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them: $F = G \cdot m_1 \cdot m_2 / r^2$, where m_1 and m_2 are the masses, r is the distance between them, and G is the universal gravitational constant, $G \approx 6.6738 \times 10^{-11} \text{ m}^3/\text{kg/s}^2$. This law is at the heart of the GRACE measurements, since any specific mass on the Earth is in general at a different distance from the two spacecrafts, causing a slight difference in the gravitational acceleration they impart to the spacecraft, and thus causing a slight but measurable relative acceleration between the spacecrafts.

Satellite Is an object, natural (like the Moon) or artificial (each GRACE satellite) that orbits around

another large object, in this case the Earth. "Orbits" means that the centripetal acceleration due to the speed of the satellite equals the gravitational acceleration between the satellite and the larger object it orbits around; in this manner the satellite neither falls toward Earth, nor escapes its gravitational pull. In practice, the GRACE satellites do fall slightly toward the Earth while they orbit around it, whereas the Moon slowly increases its distance to the Earth.

Spherical harmonics Are a set of functions of latitude and longitude that form an infinite, orthogonal, normalized set of basis functions whose sum, with appropriate scale coefficients, completely describes any other function defined in terms of spherical coordinates. Spherical harmonics satisfy Laplace's equation, as does the gravitational potential outside the Earth. Laplace's equation states that the sum of the second derivatives of the gravitational potential with respect to each of the three directions of space at a point must add up to zero if there are no masses at that point.

Definition of the Subject and Its Importance

The gravity field of the Earth, caused by the distribution of masses inside and on the surface of the Earth, changes in time due to the redistribution of mass. Such mass fluxes can be due both to natural processes (such as the seasonal water cycle, ocean dynamics, or atmospheric variations), as well as due to human actions, such as the systematic withdrawal of groundwater for human consumption. The ability to measure such changes globally is of great significance for understanding the environmental dimension of sustainability.

Until the launch of the GRACE satellite pair in 2002, such time changes in mass redistribution could only be measured globally as time changes in the longest wavelengths of the gravity field, on the order of 10,000 km and longer, from the orbit perturbations of artificial Earth satellites, or very locally at individual points on the Earth using long-term gravimeter deployments.

The GRACE satellite pair has provided the first global measurements at horizontal resolutions from 300 km to global, and time scales from 10 days to

interannual. These measurements have been used to assess the mass variability in the oceans, terrestrial water storage, and loss of ice mass in glaciers and ice sheets. Long-term trends, episodic variations such as result from large earthquakes, and seasonal changes have all been measured with unprecedented accuracy and detail.

Introduction

Imagine the ability to weigh Greenland every month over several years: It would tell us whether Greenland is losing ice, how fast, and whether the loss is accelerating. Imagine the ability to weigh the total water content every month over several years in the soil of Northern India, where the groundwater supplies hundreds of millions of people: It would tell us whether that precious resource is being used at a sustainable rate or depleted. Imagine the ability to measure how much water the ocean basins are gaining: It would allow us to separate the two key components of sea-level rise, the addition of mass and the thermal expansion due to increased ocean temperature. As we will see below, these are no longer imaginary possibilities but actual findings using data from a satellite pair called GRACE, or Gravity Recovery and Climate Experiment. The GRACE satellite pair was launched on March 17, 2002 to measure monthly changes in the gravity field of the Earth with exquisite accuracy. It is from these monthly changes in the gravitational attraction of large bodies of water, ice, and rock that we can “weigh” changes from month to month in Greenland, groundwater, and even the oceans. GRACE also yields a very accurate measurement of the time-averaged gravity field, a useful quantity for studies of tectonics and of the time-averaged ocean circulation. There is a well-known inherent ambiguity in interpreting changes in gravity: The data type can localize the changed mass horizontally, but it alone cannot distinguish whether the mass change comes from the surface or deeper layers, from the ocean, or from an earthquake that moved the sea floor. Additional information is needed to identify the source of the mass changes unambiguously. This entry reviews the basis of the GRACE measurements, some technical details that clarify the strengths and limitations of the GRACE data, and provides illustrative examples of the applications in hydrological,

cryospheric, oceanic, and geophysical sciences. Given the wide variety of phenomena GRACE can tell us about, this entry cannot be an exhaustive review of the more than 700 peer-reviewed publications on GRACE published through the end of 2010. Also, while the GRACE satellite pair carries global position systems (GPS) antennas and hardware for use in “GPS occultation” soundings of the atmosphere, we focus the discussion on results obtained from the gravity data derived from GRACE.

The measurement of gravity from satellite orbit perturbations (that is, from departures of the satellite’s path from a classical Keplerian ellipse) has been in existence for many decades. It started immediately at the dawn of the space age, with measurements of the Earth’s pear shape [62]. Tracking satellites from terrestrial observatories or from GPS satellites, using radiometric or laser ranging methods, provided important information on the long-wavelength, static (long-term mean) gravity field through the next 3 decades. Precise measurement of time-variable gravity from those techniques was only possible, however, for the Earth’s oblateness parameter (J_2). Long-term measurements of the J_2 parameter were used [24, 27] to identify and explain a steadily decreasing trend in the flattening of the Earth, due to GIA, and associate the departures from that steady trend with interannual climatic variability, such as the El Niño-Southern Oscillation phenomenon. See also [8, 22, 23]. The advent of the German CHAMP mission in mid-2000 improved the situation; CHAMP determined gravity changes at continental (several thousand kilometers) scales and annual and longer periods. The GRACE mission results provided a true paradigm shift, providing an unprecedented global and accurate measurement to 300-km spatial resolution, continuously since 2002.

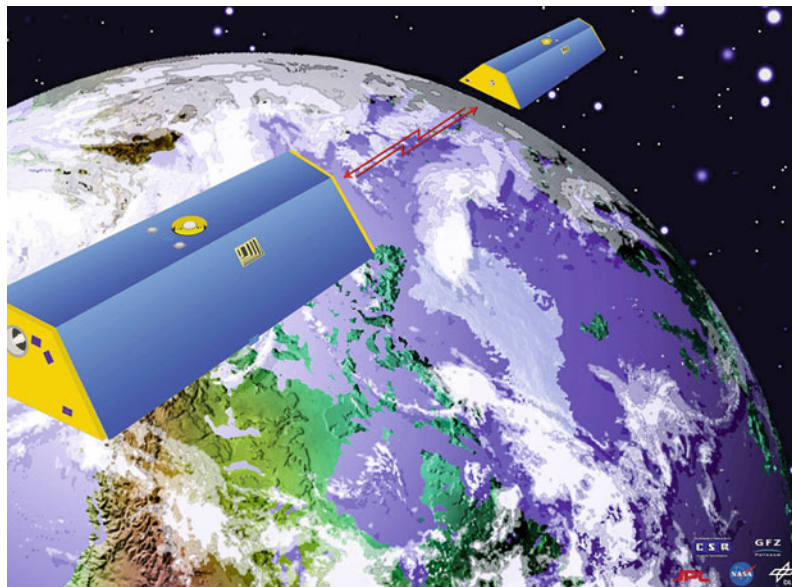
What Is GRACE

GRACE, launched on March 17, 2002, is a joint mission of the US and German space agencies, the National Aeronautics and Space Administration (NASA) and the Forschungszentrum der Bundesrepublik Deutschland für Luft- und Raumfahrt (DLR). GRACE was the first mission flown under the NASA Earth System Science Pathfinder (ESSP) Program, with Principal Investigator (PI) from the Center for Space

Research at the University of Texas, and the co-PI from the Helmholtz Center Potsdam, German Research Centre for Geosciences (GFZ) [89]. Mission management and implementation are the responsibility of the NASA Jet Propulsion Laboratory (JPL), a NASA center managed by the California Institute of Technology. Mission operations are performed by DLR, with co-funding by the European Space Agency (ESA). DLR was also responsible for the launch services. JPL led the development of the satellite system in partnership with Astrium GmbH and Space Systems Loral (SS/L). Astrium was the prime contractor for major elements of the two satellites based on the CHAMP (Challenging MiniPayload) satellite heritage, while SS/L was the prime contractor for the attitude control system and microwave instrument electronics. JPL also provided the key microwave range-rate sensors between the satellite pair, and the Global Positioning System (GPS) receivers.

Almost all Earth-viewing spaceborne instruments measure electromagnetic radiation, either naturally

emitted or reflected by the Earth's surface (passive), or the returns of signals emitted toward the Earth by the instrument itself (active). GRACE is fundamentally different: The satellite pair senses gravity through the changes in the intersatellite distance. The two satellites fly along the same orbit (Fig. 1). Currently (as of July 2011), the pair flies 454 km above the Earth's surface (at launch, it was 502 km), separated by 206 km. Imagine the satellite pair passing over a mountain. As the lead satellite "feels" the increased gravitational attraction of the upcoming mountain, it accelerates toward that mass excess. Due to Newton's law of gravitational attraction (whereby the force acting on a mass is inversely proportional to the squared distance to the attracting mass) that acceleration toward the mountain is larger for the lead satellite than for the trailing one, whose distance to the mountain is larger. As a consequence, the distance between the two satellites increases, if ever so slightly. When the mountain is in-between the satellites, both are attracted toward the mountain, so the intersatellite distance decreases.



Gravity Recovery and Climate Experiment (GRACE): Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains. Figure 1

Artist's conception of the GRACE satellite pair in orbit. In reality, each satellite is 3.123-m long, the distance between the pair ranges between approximately 175 and 230 km, and the altitude has been slowly decaying from 502 km at launch to 454 km currently (July 2011), relative to a distance of 6,378 km from the center of the Earth. The orbital inclination is 89°

A highly precise microwave ranging system (called “KBR,” or K-band ranging system) aboard each satellite beams radio signals toward the other satellite and measures those changes in distance: This measurement forms the basis of the computation of the gravitational field. The measurement is so precise that it detects changes of a few microns in range or approximately $0.1 \mu\text{m/s}$ in velocity every 5 s. To picture that level of accuracy, it is approximately equivalent to measuring the distance between Los Angeles and San Diego to an accuracy of a (fine) human hair’s width.

The key science instrument in each GRACE satellite is the microwave K-band ranging instrument. Each satellite transmits signals to the other at two frequencies, (24 and 32 GHz, called “K” and “Ka” bands, respectively) in order to correct for ionospheric path delay. The K-band ranging assembly consists of an ultra-stable oscillator, the microwave electronics, a K-band ranging horn, sampler, and the instrument processing unit (IPU). The ultra-stable oscillator serves as the frequency and clock reference for the GRACE satellites. The K-band ranging horn transmits and receives K-band carrier signals to and from the other GRACE satellite. The IPU is the nerve center for the science instruments in the spacecraft, providing the digital signal processing functions for the K and Ka band signals, as well as for the GPS signals. It also provides various clocks for the satellite operations and performs data processing for the star camera attitude (directions). The K-band ranging system was manufactured by JPL with equipment from Space Systems/Loral and the Applied Physics Laboratory.

The intersatellite range-change measurement must be corrected for the effect of forces other than gravity, such as drag or solar radiation pressure. To do so, both satellites carry accelerometers located very close to each satellite’s center of mass. The accelerometers were built by ONERA, the French Aerospace Lab. An approximate orientation of the spacecraft is provided by the star cameras (two aboard each satellite). Used both for science and for attitude and orbit control, they determine each satellite’s orientation by tracking their orientation with reference to the stars. They were developed by the Danish Technical University, Copenhagen, Denmark. In order to provide precise synchronization of the ranging data between the two satellites, and in order to geolocate the science data, both satellites carry Global

Positioning System (GPS) units developed by JPL. There are three GPS antennas: One is used to collect navigation data, the two other antennas are used for backup navigation and atmospheric occultation data collection. The GPS units provide an additional piece of information essential to the computations: time-tagging events to better than 0.1 ms.

Other instrument assemblies aboard the spacecraft include a Coarse Earth and Sun sensor, used for approximate position and orientation whenever GRACE is in “safe mode” (when a serious anomaly in the satellite cannot be automatically corrected on board, the satellite is put into “safe mode” by the onboard computer, to ensure only the satellite’s vital functions remain on, while the failure is being analyzed on the ground); a Center of Mass Trim Assembly, to adjust the satellite’s center of mass with respect to the proof mass of the accelerometers; solar cell arrays, covering the outer shell of the satellites to generate power for the electronics, and nickel-hydrogen cell batteries to store and release power as needed; telemetry and telecommand subsystems for the satellites to communicate with Earth via radio systems in the microwave S-band spectrum. Each satellite uses a separate set of S-band frequencies for transmission and reception. Both, the power and telemetry as well as the telecommand subsystems, were supplied by Astrium.

The satellite’s “attitude,” or orientation and orbit control, is controlled by a system consisting of sensors, actuators, and software. Two kinds of attitude actuators are available. A reaction control system with a set of twelve 10-mN thrusters uses gaseous nitrogen stored in the two tanks along the main satellite axis. Fine corrections of orientation are adjusted using six 30-Amp- m^2 magnetorquers, to minimize the satellite fuel consumption. Each GRACE satellite can adjust its orbit by firing its two orbit-control thrusters (also gaseous nitrogen propellant) mounted on the rear-panel of the satellite, each of which provides 40 mN of thrust. The attitude and orbit-control system was designed by Space Systems Loral and implemented by Astrium and its subcontractors.

Further information on the spacecraft can be found at http://grace.jpl.nasa.gov/files/GRACE_Press_Kit.pdf, and details on the mission, its subsystems, and current operational status are available at the GRACE mission homepage: <http://www.csr.utexas.edu/grace/>.

Spherical Harmonics; Equivalent Water Thickness; Mascons; and Spatial Resolution

The gravitational potential of the Earth, sensed by a satellite at altitude h , is formally expressed as the sum of spherical harmonic functions [8]

$$T(\theta, \varphi) = \frac{GM}{(a+h)} \sum_{l=0}^{\infty} \sum_{m=0}^l \left(\frac{a}{a+h} \right)^l (C_{lm} \cos(m\phi) + S_{lm} \sin(m\phi)) P_{lm}(\cos(\theta)) \quad (1)$$

where T is the “anomalous” potential (the difference between the true potential and a reference value) at colatitude θ , longitude ϕ . The dimensionless C_{lm} ’s and S_{lm} ’s are called Stokes’ coefficients; the P_{lm} are Legendre functions of degree l , order m , $m \leq l$, GM is the product of the universal gravitational constant G times the mass of the Earth, M , and a is a mean radius of the Earth. All other quantities (geoid height, gravity acceleration anomaly, or disturbance, etc.) have similar expansions, whose coefficients are related to those in Eq. 1 through functions of the degree l and dimensional constants. The $l = 0, m = 0$ coefficient in the expansion is exactly 1.0, and accounts for the total mass of the Earth system, conventionally regarded as a constant. The $l = 1$ terms are zero if the analysis is done in a reference frame where its origin is the instantaneous center of mass of the Earth system. The $n = 2, m = 0$ term accounts of the Earth’s oblateness, and is of order 10^{-3} relative to the central term. All other harmonic coefficients are of order 10^{-6} or smaller. Representation of the geographic features of small spatial extent requires the expansion to be carried to higher degree- l .

The GRACE data analysis problem reduces to the extraction of the Stokes coefficients from residuals of the GRACE tracking data calculated using prior best knowledge of the Earth’s gravity field and its variations. A global distribution of the mission data is required before the gravity field parameters can be estimated. It takes approximately 1 month for the GRACE data to provide uniform, global coverage. The Earth’s gravity field variations are therefore represented by monthly piece-wise constant spherical harmonic coefficients of the anomalous geopotential, represented to a fixed maximum degree/order. The GRACE mission data products

are these Stokes coefficients, delivered monthly for the entire mission lifetime. The long-term mean estimates provide the determination of the static gravity field, and the monthly deviations represent the time-variability of the mass flux at the longer time scales.

However, hydrologists, oceanographers, and cryospheric scientists are less interested in the gravitational C_{lm} ’s and S_{lm} ’s coefficients, but rather in estimates of time changes in surface mass, or ocean bottom pressure. Let the surface mass density σ be the vertical integral of the density ρ through the Earth’s surface layer (containing the atmosphere, the oceans, and the water/snow/ice stored on land), where we assume all time changes are concentrated. Using the above series, with $h = 0$, σ is

$$\Delta\sigma(\theta, \phi) \approx \frac{a\rho_E}{3} \sum_{l=0}^{l_{\max}} \sum_{m=0}^l \frac{(2l+1)}{(1+k_l)} W_l P_{lm}(\cos \theta) \times [\Delta C_{lm} \cos(m\varphi) + \Delta S_{lm} \sin(m\varphi)] \quad (2)$$

where ρ_E is the mean density of the Earth, a is its radius, and the k_l are so-called load Love numbers representing the elastic response of the solid Earth to surface loading [96]. The W_l is an isotropic filter that strongly downweights high degree l terms, which are often contaminated with noise amplified by the downward continuation effect, discussed below. More general filters, nonisotropic filters and irregular area-averaged filters, are discussed by Swenson and Wahr [84], while a decorrelation filter which removes North–South “striping” from GRACE maps was first discussed by Swenson and Wahr [85], then optimized for ocean studies by Chambers [10]. Equation 2 allows us to convert the C_{lm} , S_{lm} to a surface mass distribution, typically expressed in centimeters of equivalent water thickness (“cm EWT”). In the above formula, the maximum spherical harmonic degree indicated is l_{\max} ; in practice, there is little signal above spherical harmonic degree 60 in the monthly GRACE fields, although some solutions are given to $l_{\max} = 120$. An order of magnitude estimate for the spatial half-wavelength associated with a particular degree l is given by the approximation $40,000 \text{ km}/(2l)$ so $l = < 60$ implies half-wavelengths longer than ~ 330 .

Two sets of coefficients require special handling. The coefficients of degree 1 orders 0 and 1, which

indicate the position of the center of mass of the Earth (which varies in time) relative to an Earth-fixed coordinate origin, are not measured by GRACE. Swenson et al. [87] realized that by assuming that we know the component of $n = 1$ from an ocean model, it is possible to use the GRACE data and the ocean information to derive degree 1 coefficients. Degree 1 coefficients can also be derived from satellite laser and Doppler ranging data [25]. The coefficients of degree 2 order 0 are poorly determined in the GRACE data, and it is now a standard practice to replace them by coefficients estimated from satellite laser ranging [23]. See also [9].

Another way to describe the gravity field uses “mascons” (mass concentrations, see “Glossary”). In this approach, either small spherical caps or point masses are assumed to cover the Earth’s surface, and one solves directly for the mass of each local mascon from the intersatellite range-rate or acceleration data, rather than for the global Stokes coefficients. When the solution is unconstrained, such that no a priori correlation is imposed on neighboring mascons, nor any smoothing is imposed on the Stokes coefficients, there is little difference in the result [79]. However, there are advantages in imposing correlations between neighboring mascons when one knows that, for example, a set is in one hydrological basin and the nearby set is in another one, uncorrelated to the first. In addition, the mascon basis functions more conveniently allow higher spatial resolution at higher latitudes where ground tracks are spaced more densely.

GRACE data products can be obtained from several public sources. Because of small differences in the processing strategy, the results differ in small but significant ways, and it is a good practice to check the results from two centers for a particular application. The GRACE Science Data System consists of three centers, the Center for Space Research (UTCSR) at the University of Texas-Austin [5], the Geoforschungszentrum in Potsdam, Germany, and the Jet Propulsion Laboratory in Pasadena. These three centers provide spherical harmonic solutions through <http://podaac.jpl.nasa.gov> and a mirror archive at <http://isdc.gfz-potsdam.de/>; gridded versions of their data, with additional corrections applied (for example, the decorrelation filters of Chambers [10] or Swenson and Wahr [85], can be found at <http://grace.jpl.nasa.gov>. Additional sources include NASA’s Goddard Space

Flight Center (GSFC) [79] who supply mascon solutions, the Centre National d’Etudes Spatiales (CNES) in France [7, 46] who supply 10-day SH solutions, the Institut für Geodäsie und Geoinformation at the University of Bonn (ITG-Bonn), Germany who supply both monthly and daily SH solutions [58], and Delft University of Technology, Netherlands [49] who supply monthly SH solutions.

Background Fields

Since the satellite motion is a nonlinear function of the gravitational potential, the derivation of a monthly set of either Stokes coefficients or mascons involves the solution of a very large nonlinear least squares problem. Nongravitational influences are modeled using the data collected by the accelerometer. The satellite orientation is also important because the range-change measurements are made to a reference point displaced from the center of mass of the satellite, to which the dynamical equations of motion refer. To make the solution computationally tractable, one solves not for the full gravity field, but for the difference from an initial guess (a nominal, a priori model). The a priori model is the sum of an earlier estimate of the time-averaged gravity field, plus the disturbances due to the Sun, Moon, and other planets, plus the gravitational effects of solid Earth and ocean tides, plus the time-varying mass of the atmosphere derived from European Center for Medium Range Weather Forecast (ECMWF) model output, plus the time-varying effects of non-tidal ocean mass redistribution derived from a numerical ocean model driven by ECMWF [28], plus a model of the “pole tide” (both Earth and ocean), which is not actually a tide but the result of a perturbation of the Earth’s rotation. All the components of the a priori model are collectively called “background models,” and the solution depends to some extent on their fidelity. The equations of motion of the satellites are integrated numerically with time steps of a few seconds, the background models are interpolated to those times, the observable position of the satellites and intersatellite range rate are computed from the models and compared to the observed values, and the residual differences are then used in the least squares solutions. These residual differences reflect both the processes not modeled in the background

(e.g., land-surface hydrology, or ice sheet mass changes) as well as processes erroneously modeled in the background (e.g., atmospheric or tidal variability). The residual differences, aggregated over a month, are then used in the least squares solution for the anomalous geopotential for that month. In addition to the desired gravitational model for a month, a large number of instrument- and spacecraft-dependent “nuisance parameters” are adjusted which include biases, bias drifts, clock offsets, orbits of the GRACE satellites and the GPS satellites, etc.

Note that the tides, atmospheric and oceanic mass redistributions, as well as land hydrology, all have energy at periods much shorter than a month, the typical time span of one GRACE gravity solution. The models therefore remove this energy, which could otherwise be aliased into longer period variations in the retrieved gravity fields.

For ocean studies, the monthly averaged ocean model is added back to the solution. At this writing, adding a background model of the land hydrological mass changes, which has been shown to improve the solution, is being considered to become part of the gravity field processing for an upcoming data release.

Ambiguity; Downward Continuation; Data Accuracy

Gravity data provided by in-situ or airborne instruments (called gravimeters) have long been used in both exploration and research geophysics to help define buried structures that exhibit density variations (e.g., ore deposits). Two measurement effects are well known by those practitioners: the ambiguity of interpretation (“nonuniqueness” of a solution), and the problem of “downward continuation” of a measurement taken at some altitude.

Consider a sphere of density ρ_1 and radius r_1 at depth $h_1 > r_1$. This sphere would produce the same gravitational attraction at the surface as another sphere of density ρ_2 and the same radius r_1 but at depth h_2 , exactly “under” the first sphere, so long as they satisfied the ratios $\rho_1/\rho_2 = (h_2/h_1)^2$. While this is harder to see in more complex shapes, it illustrates well the ambiguity in the interpretation: It is not possible to assign a mass uniquely to a buried structure without some additional data or constraints [66]. The main assumption to aid

the interpretation of GRACE data is that monthly changes in gravitational attraction are produced primarily by movement of water within and among Earth reservoirs: the cryosphere, the oceans, and the soil [96], except in those regions where glacial isostatic adjustment (GIA) and earthquakes produce strong signals. This assumption is generally valid: While large mountain ranges produce large signals in the time-mean gravity field, their effect in the time-varying gravity field over a few years or even decades is negligible (Fig. 2). Indeed, as will be shown below, maps of the time-variable gravity do not show strong signals where the time-averaged gravity field has its stronger signals. In the case of GRACE, ambiguity arises when trends in mass displacements from GIA [due to the lithosphere’s slow adjustment to the loss of the ice load that peaked 21,000 calendar years ago [65] and was essentially complete 6,000 years ago] can confound the signals due to trends in present-day ice or hydrological mass changes. Large earthquakes can also cause prominent signals in GRACE data as the lithosphere moves and adjusts following a quake.

To understand downward continuation, consider a simple plane, one-dimensional geometry. The gravity field above the x -axis, due to a mass distribution below the x -axis, satisfies the following equation, a consequence of Laplace’s equation which governs the gravity field away from its source masses

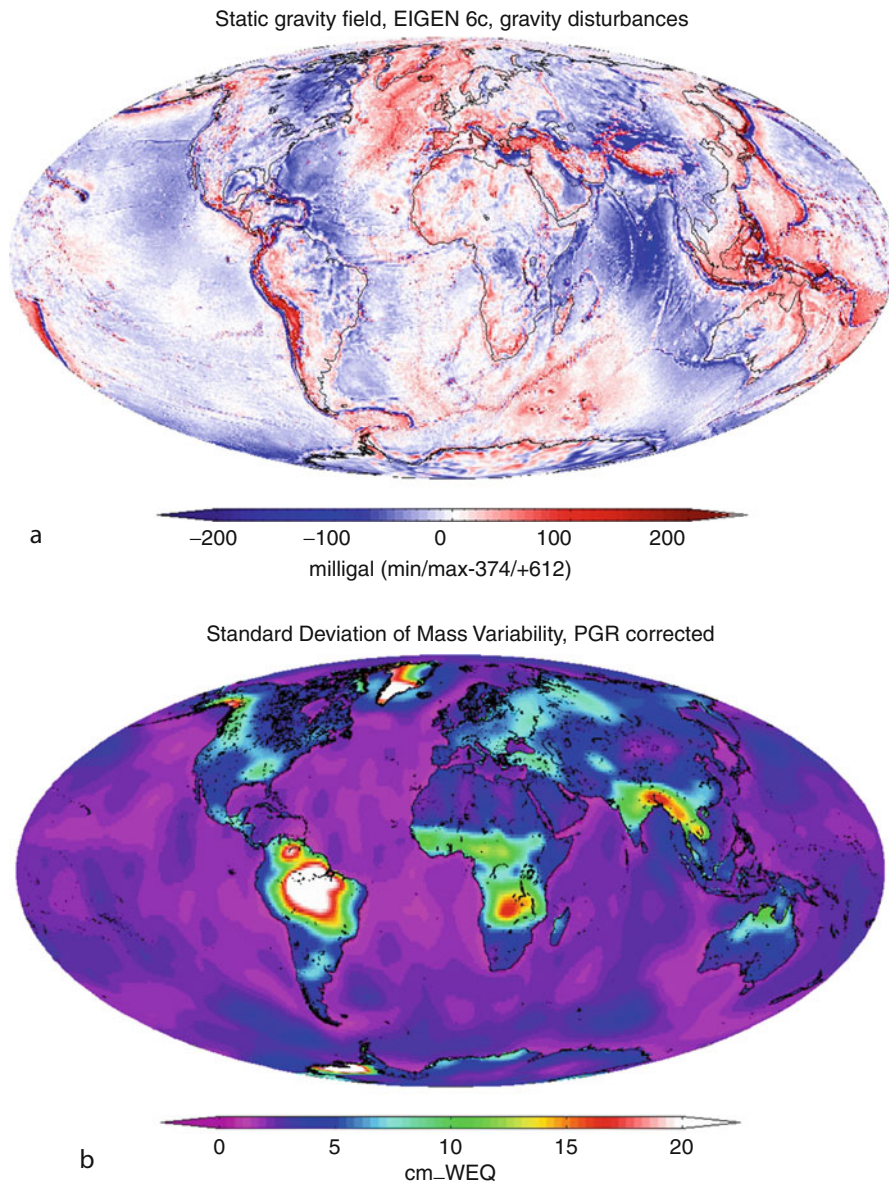
$$G(\lambda, z) = G(\lambda, 0) \exp\left(\frac{-2\pi z}{\lambda}\right) \quad (3)$$

where $G(\lambda, z)$ is the Fourier component of the gravity acceleration distribution $g(x, z)$ with wavelength λ , and z is the altitude above a plane that is (just) above all masses. The exponential in Eq. 3 is the plane equivalent of the factor $(a/(a + h))^1$ in the spherical harmonic expansion (Eq. 1). As a consequence, if we measure $g(x, z)$ at altitude z , short wavelength features of the gravity distribution $g(x, 0)$ are attenuated by the factor $\exp(-2\pi z/\lambda)$; for $\lambda \gg z$, this factor tends to 1 while for $\lambda \ll z$, the factor tends to zero. The “downward continuation” problem arises when we measure at altitude z where short wavelengths are attenuated, but wish to know the signal magnitude at $z = 0$. This would require an exponential amplification of the weak signals, but it would also produce an exponential magnification of

the inevitable noise present in the data. Filter applications to solve or at least mitigate this problem were discussed in section “[Spherical Harmonics](#), [Equivalent Water Thickness](#), [Mascons](#), [Spatial Resolution](#).”

The accuracy of GRACE data at a particular point in space and time is the result of several factors: propagation of measurement errors through the processing,

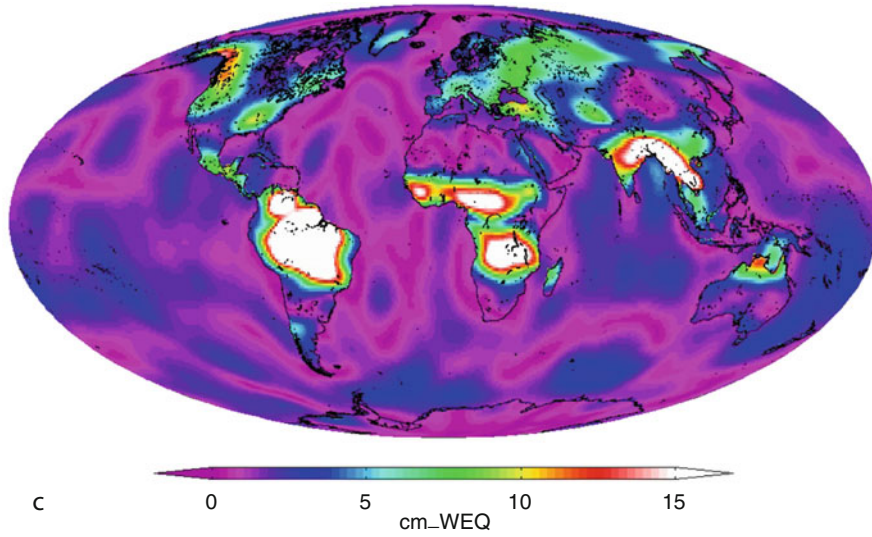
loss of signal due to smoothing or destriping which makes the values at a geographical position include “leakage” from neighboring positions (smoothing) and even far away locations (destriping), and the ambiguity associated with the signals of GIA and large earthquakes. The effect of measurement errors depends on latitude (smaller at higher latitudes), and



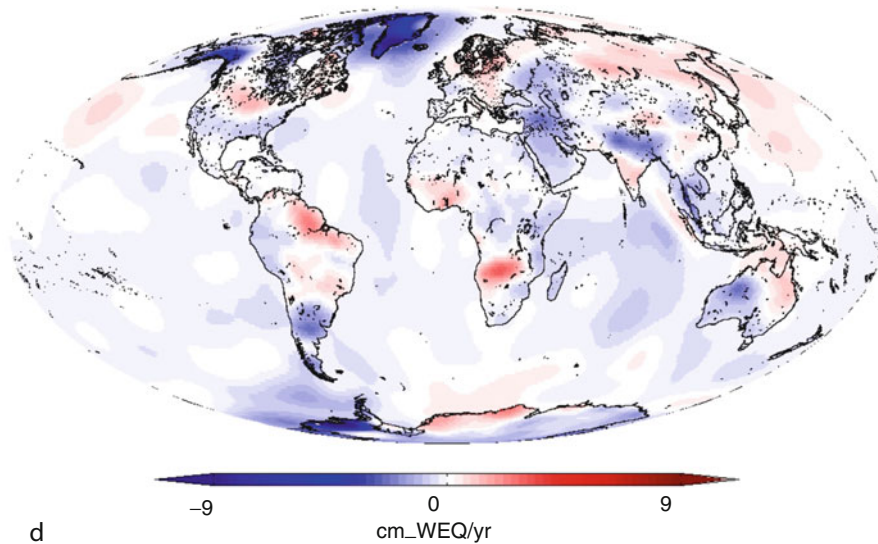
Gravity Recovery and Climate Experiment (GRACE): Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains. Figure 2

(Continued)

Annual Cycle (sin amplitude) of Mass Variability, PGR corrected



Trend of Mass Variability, PGR corrected



Gravity Recovery and Climate Experiment (GRACE): Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains. Figure 2

(a) The static gravity field EIGEN6c, which combines GRACE and in situ data. The quantity displayed is the “gravity disturbance,” which is the difference between actual gravity acceleration at a point and a reference value, which includes the longest wavelengths of gravity, at the same point. The short length-scale-features in this map are closely related to surface topography, whether on land or at the sea floor. (Data from the International Centre for Global Earth Models, Potsdam, Germany <http://icgem.gfz-potsdam.de/ICGEM>.) (b) The standard deviation of monthly maps of changes of mass which give rise to changes in the gravity field derived from GRACE, expressed in centimeters of equivalent water thickness (see section “Spherical Harmonics. Equivalent Water Thickness. Mascons. Spatial Resolution”; here abbreviated as “WEQ” = “EWT”). Data cover the time period Jan 2003 to Dec 2010. A destriping decorrelation filter has been applied; land data are further smoothed with a 300 km Gauss filter, while ocean data are smoothed with a 500 km Gauss filter. The trends due to glacial isostatic adjustment have been removed from the data using the model of (Paulson et al. [65]). Note that there

smoothing radius (smaller with larger radii); for example, smoothing with a 750 km Gaussian, but not applying any destriping filter, leads to errors in the mass anomalies ranging between 8 mm of equivalent water near the poles to over 28 mm at the Equator [97].

Applications in Land Hydrology

Over land, repeat GRACE gravity data are interpreted as changes in total water stored in groundwater, soil moisture, vegetation, surface water, snow, and ice, a vertical integration over all relevant layers; this quantity is called terrestrial water storage (TWS). The main complicating issue that arises is the effect of destriping and smoothing filters on the hydrological signals. This effect can be estimated by applying the same filters to output of a land-surface hydrology model (LSM), then computing a gain factor as a least squares ratio between the time series of change in TWS of the filtered and unfiltered model outputs [43].

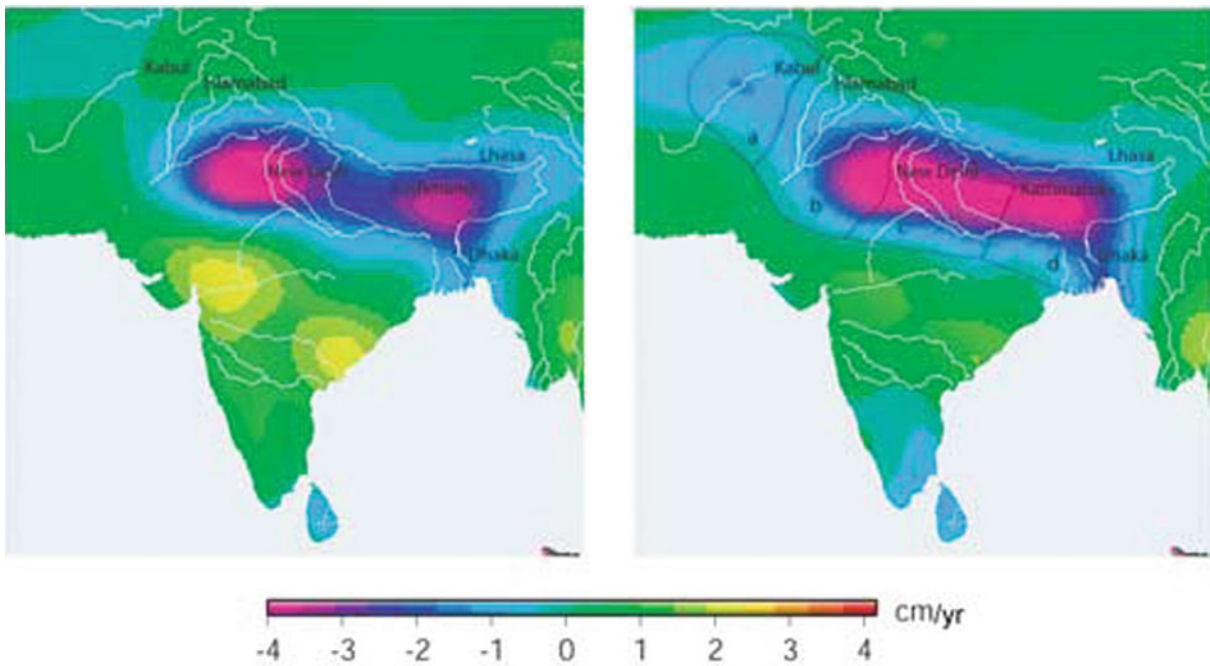
GRACE gravity data in Northern India, interpreted as TWS, revealed a steady, large-scale mass loss beyond the natural variability simulated by a land hydrology model [90], which resulted from the excessive extraction of groundwater (Fig. 3). When the GRACE data were combined with hydrological models to remove natural variability, the authors concluded the region had lost groundwater at a rate of $54 \pm 9 \text{ km}^3/\text{year}$ between April 2002 and June 2008 and noted that continued extraction of groundwater at that rate could lead to a major water crisis (see also [78]). GRACE data also showed groundwater depletion in California: During a drought period that lasted from October 2003 to March 2010, California's Sacramento and San Joaquin River Basins lost water at a rate of $31.0 \pm 2.7 \text{ mm/year}$ equivalent water height, or a total volume of 30.9 km^3 . Adding other data and a hydrological model, the authors concluded that most of the loss, 20.3 km^3 , occurred in the Central

Valley [30]. While the authors concluded that such a water loss is unsustainable over many years, the most recent data show a return to pre-drought conditions starting in 2009, likely due to significantly increased precipitation.

Australia's recent record droughts motivated many studies using GRACE data. LeBlanc et al. [45] investigated the devastating drought which started in 2001 in the Murray-Darling Basin in southeast Australia. They combined GRACE with in situ and simulated hydrological data to show that groundwater levels continued to decline 6 years after the 2001 onset of the drought, for a total groundwater loss of 104 km^3 between 2001 and 2007; the drought continued even though annual precipitation in the region returned to average during 2007. Combining GRACE data, numerical model output, and various atmospheric datasets, Garcia-Garcia (2011) [31] showed that interannual changes in the seasonal signal in TWS were intricately related both to the El Nino-Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD) phenomena, with positive phases of the IOD related to anomalously low precipitation in southeastern Australia due to a reduced tropical moisture flux, and that the sustained water storage reduction over central and southern Australia during 2006–2008 was associated with three consecutive positive IOD events. See also [3].

GRACE data over the largest drainage basin in the world, the Amazon, were used to observe two extreme seasons: the very dry 2002–2003 and the extremely wet 2009 seasons, during which (March 2009) TWS in the entire basin was $\sim 624 \pm 32 \text{ Gt}$ above the 2002–2009 time average [20]; the authors note that this huge TWS excess is roughly equal to the water consumption for 1 year in the USA. Not surprisingly, the GRACE data were consistent with precipitation data from the Global Precipitation Climatology Project. The authors also note the close correlation between these interannual TWS changes in the Amazon and El Nino-Southern

is essentially no correlation between this map and Fig. 2a; by contrast, here the largest signals are at low latitudes, West Antarctica, Greenland, and the Gulf of Alaska, indicating that by and large monthly changes in gravity are indeed associated with changes in the surface water and ice mass redistribution (Data from <http://grace.jpl.nasa.gov>). (c) The amplitude of a (sinusoidal) annual cycle fit to the data of Fig. 2b. Note the differences from Fig. 2b, for example, Antarctica where the annual cycle is almost absent, or Greenland where it is minimal. (d) Trend of the data in Fig. 2b. Note that the strongest negative (decreasing mass) trends are in Greenland, West Antarctica, and Alaska



Gravity Recovery and Climate Experiment (GRACE): Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains. Figure 3

Left: Mass loss detected by GRACE in Northern India, in centimeter of equivalent water thickness (From [90]). *Right:* after subtracting from the *left* map the prediction of the Community Land Model (CLM) hydrology model, a measure of natural variability. Thus, the *right* map depicts the anthropogenic withdrawal of groundwater in the region. See [90] for details

Oscillation (ENSO) events as measured by a widely used ENSO index based on Sea Surface Temperature in the eastern tropical Pacific. The 2002–2003 dry season was tied to 2002–2003 El Niño and the 2009 flood to La Niña conditions. See also [1, 34].

Swenson and Wahr [86] combined GRACE with radar and laser altimetry, scatterometer, and precipitation data to assess water level changes in several lakes in the Great Rift Valley, East Africa, and to separate the effects of climatic variability from those of water resources management at the downstream dams of Lake Victoria. Comparing the water level drops from Lakes Victoria, Tanganyika, and Malawi, the authors concluded that about half of the decrease in Lake Victoria was due to climate forcing, and half to water management; this conclusion for the period 2003–2009 was consistent with previous findings for earlier time periods. See also [2, 4].

An innovative application used GRACE TWS data to evaluate precipitation products from two global

analyses – Global Precipitation Climatology Project (GPCP) and Climate Prediction Center Merged Analysis of Precipitation (CMAP) – [83]. Both GPCP and CMAP merge in situ rain gauge data with satellite data. At high latitudes, the uncertainty in these products is significantly higher than at lower latitudes, due to gauge “undercatch,” a systematic underestimation by the in situ data, which is empirically corrected for. The authors used GPCP and CMAP to estimate precipitation, and the output of a land hydrology model to estimate evapotranspiration and runoff. They concluded that the gauge undercatch correction used by GPCP may be overestimated and pointed out the usefulness of the GRACE data in assessing precipitation estimates – and their biases – over large regions with sparse in situ gauge data.

The correlation of interannual variability of stored water over land and interannual changes in sea level, corrected for thermal expansion, was found to be ~ 0.6 [50, 51].

Perhaps the most promising use of GRACE data in land hydrology applications is its rigorous assimilation into numerical land-surface hydrology models (LSMs), together with a variety of other data such as snow water equivalent or soil moisture, and forcing fields such as precipitation. The reason is that the output of such models allows the separation of the various vertical components that add up to the total TWS and can provide information at scales finer than the resolution of GRACE, both in space and time. Zaitchik et al. [101] used an ensemble Kalman filter and smoother applied to the catchment land-surface model (CLSM) to study the Mississippi River basin and its four main subbasins. They found that the model with GRACE data assimilation had improved skill over the model without assimilation; skill was determined by the correlation with measured groundwater and with gauged river flow for the four subbasins and for the overall Mississippi River. They then evaluated model performance for eight smaller watersheds, all smaller than the scale of GRACE observations. In seven of eight cases, GRACE data assimilation increased the correlation between changes in TWS and gauged river flow, evidence of the potential to downscale GRACE data to finer spatial scales in hydrological applications. See also [54, 82].

Applications in Cryospheric Studies

When using the time variability GRACE data to study large glaciers and continental ice sheets, the change in gravity is interpreted as a change in ice mass; the change in equivalent water thickness, integrated over the relevant surface area and multiplied by the density of water, yields the change in Gigatons (Gt) of ice. Two complicating issues arise: glacial isostatic adjustment (GIA) and the effect of GRACE filtering and destriping on the estimates. These same issues certainly arise in hydrological and oceanographic studies, but they are more prominent in cryospheric studies for two reasons: firstly, for most hydrological studies, GIA is a small signal relative to the hydrological trend, or simply, the trend is not the main point of the study but rather interannual changes in the hydrological signal, while in cryospheric studies, the trend is the desired signal; secondly, the effect of the smoothing and destriping filters for hydrological studies can be estimated by

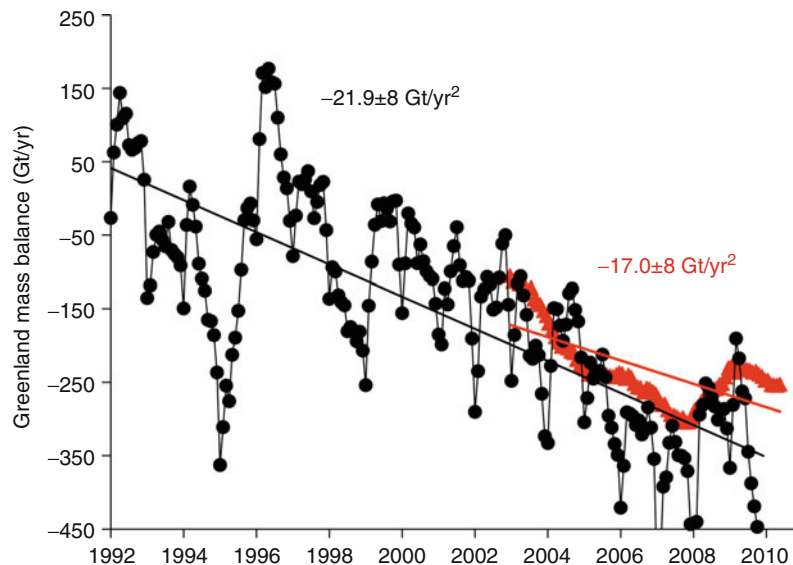
applying the same filters to output of a land-surface hydrology model (LSM). No such calculation can be performed for changes in ice mass, since currently no adequate, reliable ice sheet models exist. The GIA correction is relatively small for Greenland, but it is as large as the overall ice mass trend for Antarctica. Therefore, uncertainties in the GIA model propagate directly into uncertainties of the estimated ice loss for Antarctica. The solution to the scaling issue requires having external knowledge about the location of the small regions where ice melt is concentrated. This information could come from the Ice, Cloud, and land Elevation Satellite (ICESat) laser altimeter, or from Interferometric Synthetic Aperture Radar (InSAR), then building a model of spatial ice mass loss based on this information, and finally computing a scaling factor in a similar manner as for hydrological signals. For both the Greenland and Antarctic ice sheets, the mass loss is concentrated in coastal outlet glaciers and ice streams, and this information is used to construct a model of ice loss to which the destriping and smoothing filters are applied to derive necessary gain factors. Needless to say, “whether ice sheets and glaciers are melting, and whether such melting is accelerating” are topics of great scientific and social interest.

GRACE data revealed that Greenland’s ice lost mass at an accelerated pace: While 137 Gt/year of ice were lost in 2002–2003, the rate increased to 286 Gt/year in 2006–2009, an acceleration of -30 ± 11 Gt/year² in 2002–2009 ([93]; see also [17]). The same study estimated the mass loss from Antarctica to have accelerated from 104 Gt/year in 2002–2006 to 246 Gt/year in 2006–2009, an acceleration of -26 ± 14 Gt/year². It is important to note that, while the trends are affected by the accuracy of GIA estimates, the accelerations are not because the GIA process is a trend without notable acceleration over the time periods considered. The Antarctic ice loss is mostly concentrated in West Antarctica. These results were bolstered by another study, which covered the time period 1992–2009 [75], and combined GRACE data over 2002–2009 with the “mass-budget method” (MBM) over the longer time span. In the MBM, Interferometric Synthetic Aperture Radar is used to compute glacier velocities while radio echo sounding is used to compute ice thickness; from this combination, perimeter loss is derived, and the perimeter loss is then differenced from net

accumulation computed from the sum of snowfall minus surface ablation reconstructed from a regional atmospheric model. During the common time period of 2002.9–2009.5, the GRACE data and the MBM results agree within their error bars (Fig. 4). For Greenland, the mass losses estimated from MBM and GRACE differ by ± 20 Gt/year, within their respective errors of ± 51 Gt/year and ± 33 Gt/year, and the accelerations in mass loss agree even better: 19.3 ± 4 Gt/year² with MBM and 17.0 ± 8 Gt/year² with GRACE. For Antarctica, the mass loss rates differ by ± 50 Gt/year, within the error bar of ± 150 Gt/year for MBM and ± 75 Gt/year for GRACE, while the accelerations are 13.2 ± 10 Gt/year² with GRACE data and 15.1 ± 12 Gt/year² with MBM (the 18-year MBM estimates of the accelerations are 21.9 ± 1 Gt/year² for Greenland and 14.5 ± 2 Gt/year² for Antarctica, indicating that the GRACE period is probably too short to estimate long-term accelerations due to strong interannual variability). This study estimates that during 2006, the Antarctic mass loss using the MBM was 200 ± 150 Gt/year, comparable

to Greenland's 250 ± 40 Gt/year. The total contribution from both ice sheets amounted to 1.3 ± 0.4 mm/year sea-level rise during 2006. See also [15, 17, 19, 38, 77, 80, 92]. These studies show that most of the Antarctic ice mass loss is concentrated on the West Antarctic Ice Sheet, especially the Peninsula, and that East Antarctica has experienced a small but measurable mass increase during the GRACE years.

Focusing on smaller regions, GRACE data showed that the Greenland ice loss was largest in the southeast at the beginning of the GRACE period, but in the last few years the rates have decreased in this region, while they increased in the northwest [21, 40]. The Canadian Arctic Archipelago, off the northwestern shore of Greenland, was studied using GRACE and two other independent approaches: surface mass-budget modeling plus an estimate of ice discharge and repeat satellite laser altimetry from ICESat [32]. It was found that the three approaches gave comparable results: Between the periods 2004–2006 and 2007–2009, the rate of mass loss increased from 31 ± 8 Gt/year to 92 ± 12 Gt/year,



Gravity Recovery and Climate Experiment (GRACE): Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains. Figure 4

Decreasing trends in Greenland mass, obtained from two methods: Mass Balance method (*black circles*) and GRACE (*red triangles*). See text in section “[Applications in Cryospheric Studies](#)” for details. Note the excellent agreement of these two totally independent methods between 2003 and 2010. Because this plot depicts changes in the trend, the GRACE result is unaffected by uncertainties in modeling GIA, a constant trend (From E. Rignot, 2011, personal communication based on [75])

although the authors recognize that the time series is rather short to establish reliable long-term trends. On even smaller spatial scales, ice mass loss bordering the Gulf of Alaska and northwestern Canada was studied [54] using a mascon approach to localize GRACE signals better. It was found that although there was an overall $-84 \pm 5 \text{ Gt/year}$ mass loss ($0.23 \pm 0.01 \text{ mm/year}$ in equivalent sea-level rise) between April 2003 and September 2007, the spatial and temporal variabilities were very large, and choosing another time period (April 2003 to September 2006) gave a very different trend value. See also [16, 19]. Overall, the variability observed with GRACE agrees well with regional patterns of glacier mass loss determined from aircraft altimetry and in situ data. On even smaller spatial scales, large ice mass losses in the Patagonia ice field were documented [18, 39], the latter using a mascon approach. See also [56].

Permafrost regions also received attention. Landerer et al. [44] studied the Eurasian pan-Arctic region during 2003–2009 and concluded that changes in discharge were not due to melting of excess ground ice, but rather to changes in precipitation. However, increases of water storage, while driven by precipitation, were partially co-located with regions of discontinuous permafrost, which has warmed significantly over the last decades and changes in the terrestrial hydrological dynamics are thus likely.

Applications in Ocean Studies

GRACE data for ocean studies have been interpreted in three ways: as the time-averaged geoid, as the time-varying but globally averaged total mass of the oceans, and as the time-varying spatial pattern of ocean bottom pressure (OBP). In many cases, the data have been used together with satellite radar altimetric measurements of sea surface height (SSH). In both cases, the signals are described as changes in centimeters of water height, since 1 millibar of OBP is approximately equal to the weight of 1 cm of water.

Ocean dynamic topography is the difference between an SSH map from radar altimetry and an estimate of the geoid, the equipotential surface of the Earth's gravity field that best approximates mean sea level. Slopes in the dynamic topography imply gravitational forces acting on the ocean, forces that drive

surface ocean currents. A mean dynamic topography (MDT) is a multi-year time-averaged dynamic topography, and reflects the stationary component of the surface ocean circulation. GRACE data improved our knowledge about the geoid enormously relative to preexisting models. MDTs based on SSH and GRACE only [94] were used to study the double celled gyre of the South Atlantic ocean. In combination with surface drifter and other in situ data, as well as dynamic constraints imposed by the ocean surface momentum balance [57], these new MDTs were also used to study small scale zonal striations in the ocean; GRACE constrained the longer wavelengths and the drifter data the finer space scales. See also [76]. These MDTs are now being improved with the use of data from the Gravity field and steady-state Ocean Circulation Explorer (GOCE) [42].

Time changes in the total mass of the ocean, and their relation to land-ice melt, are a topic of great climatic and societal interest. SSH increases measured by radar altimeters reflect both increases in ocean mass and thermal expansion due to increased temperatures. In addition to radar altimetry and GRACE, an in situ observing system using Argo floats (www.argo.ucsd.edu) has been measuring temperature and salinity in the upper 1,000–2,000 m of the ocean, with good coverage since 2005. In principle, it is possible to derive the thermal expansion component from Argo, and therefore, the sum of the mass component from GRACE and the thermal expansion component from Argo should add up to the altimetrically observed SSH (assuming that the current thermal expansion is small below 2,000 m). In practice, each observing system has its strengths and weaknesses (for example, GRACE data near land tend to be contaminated by large land hydrological signals; Argo coverage prior to 2005 is spotty). However, the three datasets now agree within error estimates, both on seasonal time scales and in terms of their trend [47, 48]. In fact, a discrepancy in earlier estimates of this balance [53] helped identify a problem in the in situ data [98]. See also [12, 14, 74]. The interannual variability in the spatial distribution of GRACE-derived oceanic mass change and that derived from altimetry minus Argo has also been studied [51].

GRACE data over the oceans have been compared with in situ bottom pressure recorders (BPR). BPR sites are few relative to the vastness of the oceans, they are

seldom occupied for longer than a year or two without changing instruments, and the instruments drift, especially when first installed at the bottom of the ocean. Hence, the comparisons have focused on subannual signals. The general conclusion is that there is good agreement at mid-to-high latitudes where the signal exceeds 1 cm RMS, and poor agreement at tropical and lower latitudes, where the bottom pressure signal is weaker and GRACE errors are larger [55, 59, 64, 73].

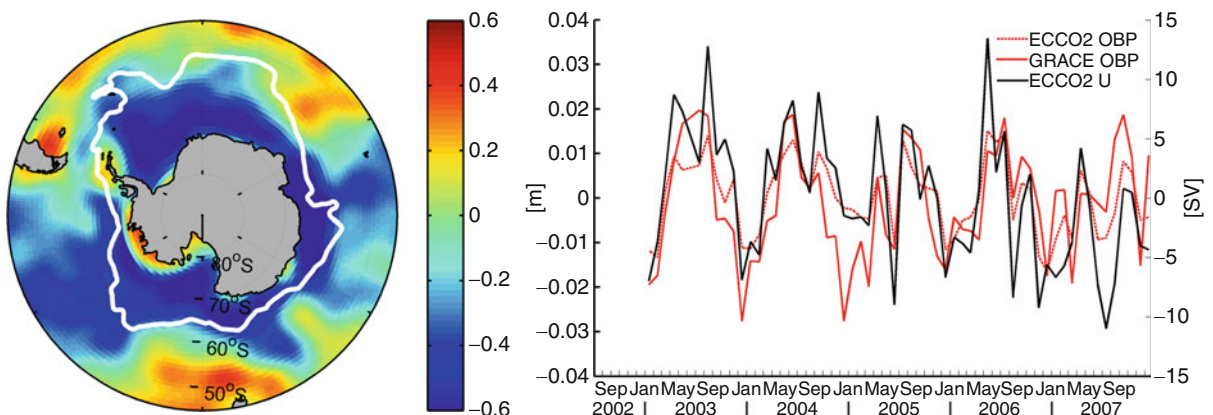
Large-scale mass exchanges among the Pacific, Atlantic, and Indian Ocean basins between August 2002 and December 2008 were identified in GRACE data and a numerical ocean model [11]. These exchanges occurred both over seasonal and interannual time scales. Changes in transport of the Antarctic Circumpolar Current were observed through changes in OBP in both GRACE and numerical ocean models [102] and (Fig. 5); an exchange of water between the Southern Ocean and the Pacific was also observed [70].

A declining trend in GRACE-derived OBP in the Arctic was correlated to corresponding changes observed with an in situ bottom pressure gauge and, more importantly, to mass changes due to decreasing upper ocean salinities near the North Pole and in the Makarov Basin [59]. (Data subsequent to the publication of that paper has shown the basin reverting to its

previous state.) In addition to this interannual variability, annual oscillations of about 2 cm OBP in the Arctic ocean, with maximum in late summer to early fall, were observed with GRACE and modeled as a response to runoff and precipitation minus evaporation that agrees in phase with the data and is 10% larger [69].

A record increase in OBP (from GRACE) and SSH (from the Jason-2 radar altimeter) was observed in late 2009 to early 2010 over a large mid-latitude region of the South East Pacific, diagnosed as a response to wind stress curl associated with a strong and persistent anticyclone in late 2009, which was likely related to the concurrent Central Pacific El Nino [6]. While all previously described results focused on large regions, a 20 cm seasonal signal in the small Gulf of Thailand was observed with GRACE, altimetric SSH and a nearby tide gage [99].

A promising use of the time-varying spatial distribution of OBP from GRACE is its incorporation into numerical ocean models as a data constraint. Steps in this direction were taken by Siegmund et al. [81] and Quinn and Ponte [71] who compared GRACE with OBP in numerical ocean models in order to assess the errors that should be assigned to the GRACE data in order to assimilate it into a numerical ocean model.



Gravity Recovery and Climate Experiment (GRACE): Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains. Figure 5

Left: Correlation of Drake Passage Transport in the ECCO2 numerical ocean model with GRACE-derived ocean bottom pressure (OBP). The white line indicates the Southern ACC front of the Antarctic Circumpolar Current. *Right:* Time series of Drake transport and OBP averaged along the Southern ACC front, both from ECCO2 and GRACE. Note that the correlation is not only over seasonal time scales but also the interannual variability is captured (Data and plots from Carmen Boening, 2010, personal communication)

The availability of GRACE data made it possible to reevaluate the impact of self-attraction and loading produced by the global redistribution of seawater and land water, as well as in the analysis of tide gage data, which typically disregard these effects (tidal models do include self-attraction and loading) [76]. The authors used continental water mass storage from a hydrological model, Greenland and Antarctica seasonal signals from GRACE, and ocean bottom pressure from a numerical ocean model, and found that the amplitude of the seasonal cycle due to self-attraction and loading at tide gages ranges between 2 and 18 mm. See also [95].

Tides are an area of special interest both as a dealiasing correction to GRACE and radar altimetry data, and more importantly because they have climatic effects; for example, they affect outflows of Antarctic Bottom Water in the Ross Sea by increasing the benthic layer thickness during spring tide [63]. Four global tidal models were assessed in terms of their ability to reduce tidal residuals (identified by their frequency) in the GRACE intersatellite range-rate data [72]; aside from power at the solar semidiurnal tide frequency in low latitudes due to errors in the model of atmospheric tides, the authors found power at the frequencies of nonlinear shallow-water tides indicating areas of needed improvement in global ocean tidal models. Going one step further, Egbert et al. [29] assimilated mass anomalies inferred from GRACE into a hydrodynamic tidal model around Antarctica focused on the M₂, S₂, and O₁ constituents. They showed that after GRACE data assimilation, the model better matched independent tide gage and ICESat laser altimetry data over the Filchner, Ronne and Ross Ice Shelves. See also [36].

Applications in Solid Earth Studies

GRACE time-variable gravity data have been applied in two areas of solid earth science: glacial isostatic adjustment (GIA, also known as postglacial rebound), and large earthquakes. The data processing for these applications is rather different from the previous examples that focus on surface mass changes. In the case of GIA, spatial trends in gravity potential as observed by GRACE are combined with in situ GPS data, ICESat laser altimetry and other data in least squares

inversions to separate the contributions to observed surface deformations caused by GIA from those caused by present-day ice mass changes. In the case of large earthquakes, the preferred data type have been the along-track intersatellite range rates, because these can better localize signals with small spatial extent that occur over a short time period.

Tide-gauge, GPS, and GRACE data were combined in a simultaneous inversion for a self-consistent model of GIA and regional sea-level change estimates for the Fennoscandia region [37]; the final models were consistent in peak uplift values (9.5 ± 0.4 mm/year), but located the peak uplift several degrees to the east of previously published results toward the middle of the northern Gulf of Bothnia; this work identified a background uniform gravity rate required for simultaneously fitting the data which could be due either to errors in the GRACE data or to errors in other models used as part of the inversion. A regional study of the Antarctic Peninsula using GRACE mascon and GPS data [38] found that simultaneous solution for ice loss and GIA crustal motions is possible, provided the linear trends in crustal uplift are robustly determined by the GPS stations with adequate spatial coverage of the regions with active GIA uplift and ice loss. A global study [100] combined GRACE and GPS data with ocean bottom pressure from a data-assimilative numerical ocean model to separate the contributions of present-day change (in terms of the thickness of ice and water) from that of GIA; these authors find a measurable GIA uplift signal in Greenland, where GIA models predict a negligible signal. Wu et al. [100] have revised their published estimate of Greenland ice mass loss from 104 Gt/year in their original paper to 140 Gt/year by adding to the inversion more realistic constraints on the ice distribution (X. Wu, 2011, personal communication). See also [13, 41, 68, 91].

Three large earthquakes were studied with GRACE data: (1) the 2004 Sumatra-Andaman earthquake with magnitude M_w 9.2, (2) the somewhat weaker 2010 Maule (Chile) earthquake (M_w 8.8), and (3) the 2011 earthquake off the Pacific coast of Tohoku, Japan (M_w 9.0). At this writing, published results are available for the first two.

The time series (May 2003 to April 2007) of localized gravity changes derived directly from the intersatellite range-rate data, using so-called Slepian

functions every 15 days, allowed [33] to find step-like time behavior (coseismic) and exponential-like time behavior (postseismic relaxation) stemming from the 26 December 2004 Sumatra-Andaman earthquake. The authors used seismic and geodetic data to estimate the coseismic slip and evaluated postseismic relaxation mechanisms that fit the GRACE data with alternate asthenosphere viscosity models. They also observed a prominent positive post-earthquake gravity change around the Nicobar Islands, attributable to seafloor uplift. See also [26]. Small changes in the GRACE satellites' intersatellite range were detected after the Maule, Chile earthquake (Mw 8.8, about 4 times lower energy release than the Sumatra-Andaman earthquake) on February 27, 2010 [35]. A gravity anomaly of $-5 \mu\text{Gal}$ over 500 km was found east of the epicenter after the earthquake. Based on coseismic models, the long-wavelength negative gravity change is primarily the result of crustal dilatation as well as surface subsidence in the onland region. Finite fault coseismic models predicted a much smaller offshore positive gravity anomaly, due to partial compensation of the gravity changes because of surface uplift and interior deformation. The authors noted that gravity data from GRACE fill in the seldom-observed long-wavelength part of the spectrum of earthquake deformations (for large earthquakes), a complement to surface geodetic measurements and seismic data.

Future Directions

The GRACE mission successfully demonstrated the usefulness of space-based time-varying gravity data in a variety of applications. It is clear that time-varying gravity data can monitor some of the effects and consequences of climate change and anthropogenic activity (ice loss; groundwater loss, the mass component of sea-level rise). Given this utility, it is not surprising that there is much interest in Europe, the USA, and in other countries to launch future missions devoted to measuring time-variable gravity. A mission called e-Motion was proposed to the European Space Agency, but was not selected in 2011. In the USA and Germany, a "follow-on" or "gap-filler" mission, very similar to the first-generation GRACE satellites, has been approved and is scheduled to launch in 2016. Several

improvements due to lessons learned from GRACE, as well as an experimental laser link between the two satellites will be implemented. In the USA, the National Research Council's Decadal Survey of Earth Sciences (NRC 2007) recommended a GRACE-II mission, which would be a significant improvement to the GRACE capabilities, and is planned to launch in the 2020 decade. Much discussion for the next decade centers around launching coordinated pairs of satellites, sponsored by different countries, to improve the spatial and temporal resolution of the retrieved signals.

Acknowledgments

This work was performed in part at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration, and at the Center for Space Research, University of Texas-Austin. Copyright 2011 California Institute of Technology.

Bibliography

Primary Literature

1. Alsdorf D, Han S-C, Bates P, Melack J (2010) Seasonal water storage on the Amazon floodplain measured from satellites. *Remote Sens Environ* 114:2448–2456
2. Awange JL, Sharifi M, Ogonda G, Wickert J, Grafarend EW, Omulo M (2007) The falling Lake Victoria water levels: GRACE, TRIMM and CHAMP satellite analysis of the lake basin. *Water Resources Manage*. doi:10.1007/s11269-007-9191-y
3. Awange JL, Sharifi MA, Baur O, Keller W, Featherstone WE, Kuhn M (2009) GRACE hydrological monitoring of Australia: current limitations and future prospects. *J Spat Sci* 54:23–36
4. Becker M, Llovel W, Cazenave A, Guentner A, Cretaux J-F (2010) Recent hydrological behavior of the East African great lakes region inferred from GRACE, satellite altimetry and rainfall observations. *C R Geosci* 342:223–233. <http://dx.doi.org/10.1016/j.crte.2009.12.010>
5. Bettadpur S (2007) CSR Level-2 processing standards document for product release 04 GRACE. The GRACE Project. Center for Space Research, University of Texas at Austin, pp 327–742. <http://podaac.jpl.nasa.gov/gravity/grace>
6. Boening C, Lee T, Zlotnicki V (2011) A record high ocean bottom pressure in the South Pacific observed by GRACE, *J Geophys Res Lett* 38:L04602. doi:10.1029/2010GL046013
7. Bruinsma S, Lemoine J-M, Biancale R, Vale's N (2010) CNES/GRGS 10-day gravity field models (release 2) and their evaluation. *Adv Space Res* 45:587–601. doi:10.1016/j.asr.2009.10.012

8. Cazenave A, Mercier F, Bouille F, Lemoine J-M (1999) Global-scale interactions between the solid Earth and its fluid envelopes at the seasonal time scale. *Earth Planet Sci Lett* 171:549–559
9. Chambers DP (2006a) Observing seasonal steric sea level variations with GRACE and satellite altimetry. *J Geophys Res* 111 (C3). doi:10.1029/2005JC002914
10. Chambers DP (2006) Evaluation of new GRACE time-variable gravity data over the ocean. *Geophys Res Lett* 33(17)
11. Chambers DP, Willis JK (2009) Low-frequency exchange of mass between ocean basins. *J Geophys Res* 114:C11008. doi:10.1029/2009JC005518
12. Chambers DP, Wahr J, Nerem RS (2004) Preliminary observations of global ocean mass variations with GRACE. *Geophys Res Lett* 31:L13310. doi:10.1029/2004GL020461
13. Chambers DP, Wahr J, Tamisiea ME, Nerem RS (2010) Ocean mass from GRACE and glacial isostatic adjustment. *J Geophys Res* 115:B11415. doi:10.1029/2010JB007530
14. Chen JL, Wilson CR, Tapley BD, Famiglietti JS, Rodell M (2005) Seasonal global mean sea level change from satellite altimeter, GRACE, and geophysical models. *J Geodesy* 79:532–539. doi:10.1007/s00190-005-0005-9
15. Chen JL, Wilson CR, Tapley BD (2006) Satellite gravity measurements confirm accelerated melting of Greenland ice sheet. *Science* 313. doi:10.1126/science.1129007
16. Chen JL, Tapley BD, Wilson CR (2006) Alaskan mountain glacial melting observed by satellite gravimetry. *Earth Planet Sci Lett* 248:353–363
17. Chen JL, Wilson CR, Tapley BD (2006) Satellite gravity measurements confirm accelerated melting of Greenland ice sheet. *Science* 313:1958–1960
18. Chen JL, Wilson CR, Tapley BD, Blankenship DD, Ivins E (2007) Patagonia icefield melting observed by GRACE. *Geophys Res Lett* 34(22):L22501. doi:10.1029/2007GL031871
19. Chen JL, Wilson CR, Blankenship DD, Tapley BD (2009) Accelerated Antarctic ice loss from satellite gravity measurements. *Nat Geosci* 2:859–862. doi:10.1038/NGEO694
20. Chen JL, Wilson CR, Tapley BD (2010) The 2009 exceptional Amazon flood and interannual terrestrial water storage change observed by GRACE. *Water Resour Res* 46:W12526. doi:10.1029/2010WR009383
21. Chen JL, Wilson CR, Tapley BD (2011) Interannual variability of Greenland ice losses from satellite gravimetry. *J Geophys Res-Solid Earth* 116:B07406. <http://dx.doi.org/10.1029/2010JB007789>
22. Cheng M, Tapley B (1999) Seasonal variations in low degree zonal harmonics of the Earth's gravity field from satellite laser ranging observations. *J Geophys Res* 104(B2):2667–2681
23. Cheng MK, Tapley BD (2004) Variations in the Earth's oblateness during the past 28 years. *J Geophys Res* 109:B09402. doi:10.1029/2004JB003028
24. Cox CM, Chao BF (2002) Detection of a large-scale mass redistribution in the terrestrial system since 1998. *Science* 297:831–833
25. Cretaux J-F, Soudarin L, Davidson FJM, Gennero M-C, Berge-Nguyen M, Cazenave A (2002) Seasonal and interannual geocenter motion from SLR and DORIS measurements: comparison with surface loading data. *J Geophys Res* 107. doi:10.1029/2002JB001820
26. de Linage C, Rivera L, Hinderer J, Boy J-P, Rochester Y, Lambrotte S, Biancale R (2009) Separation of coseismic and postseismic gravity changes for the 2004 Sumatra-Andaman earthquake from 4.6 yr of GRACE observations and modelling of the coseismic change by normal-modes summation. *Geophys J Int* 176:695–714
27. Dickey JO, Marcus SL, deViron O, Fukumori I (2002) Recent Earth oblateness variations: unraveling climate and postglacial rebound effects. *Science* 298:1975–1977
28. Dotslaw H, Thomas M (2007) Simulation and observation of global ocean mass anomalies. *J Geophys Res* 112:C05040. doi:10.1029/2006JC004035
29. Egbert GD, Erofeeva SY, Han S-C, Luthcke SB, Ray RD (2009) Assimilation of GRACE tide solutions into a numerical hydrodynamic inverse model. *Geophys Res Lett* 36:L20609. doi:10.1029/2009GL040376
30. Famiglietti JS, Lo M, Ho SL, Bethune J, Anderson KJ, Syed TH, Swenson SC, de Linage CR, Rodell M (2011) Satellites measure recent rates of groundwater depletion in California's central valley. *Geophys Res Lett* 38:L03403. doi:10.1029/2010GL046442
31. Garcia-Garcia D, Ummenhofer CC, Zlotnicki V (2011) Australian water mass variations from GRACE data linked to Indo-Pacific climate variability. *Remote Sens Environ* 115:2175–2183. <http://dx.doi.org/10.1016/j.rse.2011.04.007>
32. Gardner AS, Moholdt G, Wouters B, Wolken GJ, Burgess DO, Sharp MJ, Cogley JG, Braun C, Labine C (2011) Sharply increased mass loss from glaciers and ice caps in the Canadian Arctic Archipelago. *Nature* 473:357–360. doi:10.1038/nature10089
33. Han S-C, Sauber J, Luthcke SB, Ji C, Pollitz FF (2008) Implications of postseismic gravity change following the great 2004 Sumatra-Andaman earthquake from the regional harmonic analysis of GRACE intersatellite tracking data. *J Geophys Res* 113:B11413. doi:10.1029/2008JB005705
34. Han S-C, Kim H, Yeo I-Y, Yeh P, Oki T, Seo K-W, Alsdorf D, Luthcke SB (2009) Dynamics of surface water storage in the Amazon inferred from measurements of inter-satellite distance change. *Geophys Res Lett* 36:L09403. <http://dx.doi.org/10.1029/2009GL037910>
35. Han S-C, Sauber J, Luthcke S (2010) Regional gravity decrease after the 2010 Maule (Chile) earthquake indicates large-scale mass redistribution. *Geophys Res Lett* 37:L23307. doi:10.1029/2010GL045449
36. Han S-C, Ray RD, Luthcke SB (2010) One centimeter-level observations of diurnal ocean tides from global monthly mean time-variable gravity fields. *J Geodyn* 84:715–729. doi:10.1007/s00190-010-0405-3
37. Hill EM, Davis JL, Tamisiea ME, Lidberg M (2010) Combination of geodetic observations and models for glacial isostatic adjustment fields in Fennoscandia. *J Geophys Res* 115: B07403. doi:10.1029/2009JB006967

38. Horwath M, Dietrich R (2009) Signal and error in mass change inferences from GRACE: the case of Antarctica. *Geophys J Int* 177:849–864. doi:10.1111/j.1365-246X.2009.04139.x
39. Ivins ER, Watkins MM, Yuan D-N, Dietrich R, Casassa G, Rike A (2011) On-land ice loss and glacial isostatic adjustment at the Drake Passage: 2003–2009. *J Geophys Res* 116:B02403. doi:10.1029/2010JB007607
40. Khan SA, Wahr J, Bevis M, Velicogna I, Kendrick E (2010) Spread of ice mass loss into northwest Greenland observed by GRACE and GPS. *Geophys Res Lett* 37:6501. doi:10.1029/2010GL042460
41. King MA, Altamimi Z, Boehm J, Bos M, Dach R et al (2010) Improved constraints on models of glacial isostatic adjustment: a review of the contribution of ground-based geodetic observations. *Surv Geophys* 31:465–507. <http://dx.doi.org/10.1007/s10712-010-9100-4>
42. Knudsen P, Bingham R, Andersen O, Rio M-H (2011) A global mean dynamic topography and ocean circulation estimation using a preliminary GOCE gravity model. *J Geodesy*. doi:10.1007/s00190-011-0485-8
43. Landerer FW, Swenson SC (2011) Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resour Res*, in press
44. Landerer FW, Dickey JO, Güntner A (2010) Terrestrial water budget of the Eurasian pan-Arctic from GRACE satellite measurements during 2003–2009. *J Geophys Res* 115:D23115. doi:10.1029/2010JD014584
45. Leblanc MJ, Tregoning P, Ramillien G, Tweed SO, Fakes A (2009) Basin-scale, integrated observations of the early 21st century multiyear drought in southeast Australia. *Water Resour Res* 45:W04408. doi:10.1029/2008WR007333
46. Lemoine JM, Bruinsma S, Loyer S, Biancale R, Marty JC, Perosanz F, Balmino G (2007) Temporal gravity field models inferred from GRACE data. *Adv Space Res* 39(10): 1620–1629
47. Leuliette EW, Miller L (2009) Closing the sea level rise budget with altimetry, Argo, and GRACE. *Geophys Res Lett* 36:L04608. doi:10.1029/2008GL036010
48. Leuliette EW, Willis JK (2011) Balancing the sea level budget. *Oceanography* 24:122–129. <http://dx.doi.org/10.5670/oceanog.2011.32>
49. Liu X, Ditmar P, Siemes C, Slobbe DC, Revtova E, Klees R, Riva R, Zhao Q (2010) DEOS Mass Transport model (DMT-1) based on GRACE satellite data: methodology and validation. *Geophys J Int* 181:769–788. doi:10.1111/j.1365-246X.2010.04533.x
50. Llovel W, Becker M, Cazenave A, Cretaux JF, Ramillien G (2010) Global land water storage change from GRACE over 2002–2009; inference on sea level. *CR Geosciences* 342:179–188. <http://dx.doi.org/10.1016/j.crte.2009.12.004>
51. Llovel W, Guinehut S, Cazenave A (2010) Regional and interannual variability in sea level over 2002–2009 based on satellite altimetry Argo float data and GRACE ocean mass. *Ocean Dyn* 60:1193–1204. doi:10.1007/s10236-010-0324-0
52. Lo M-H, Famiglietti JS, Yeh P-J-F, Syed TH (2010) Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data. *Water Resources Res* 46:W0551. <http://dx.doi.org/10.1029/2009WR007855>
53. Lombard A, Garcia D, Ramillien G, Cazenave A, Biancale R, Lemoine JM, Flechtner F, Schmidt R, Ishii M (2007) Estimation of steric sea level variations from combined GRACE and Jason-1 data. *Earth Planet Sci Lett* 254:194–202. doi:10.1016/j.epsl.2006.11.035
54. Luthcke SB, Arendt AA, Rowlands DD, McCarthy JJ, Larsen CF (2008) Recent glacier mass changes in the Gulf of Alaska region from GRACE mascon solutions. *J Glaciol* 54(188):767–777. <http://dx.doi.org/10.3189/002214308787779933>
55. Macrander A, Böning C, Boebel O, Schröter J (2010) Validation of GRACE gravity fields by in-situ data of ocean bottom pressure. In: Flechtner F, Gruber T, Güntner A, Manda M, Rothacher M, Schöne T, Wickert J (eds) *System Earth via geodetic-geophysical space techniques*. Springer, Berlin. http://dx.doi.org/10.1007/978-3-642-10228-8_14
56. Matsuo K, Heki K (2010) Time-variable ice loss in Asian high mountains from satellite gravimetry. *Earth Planet Sci Lett* 290:30–36. doi:10.1016/j.epsl.2009.11.053
57. Maximenko N, Niiler P, Rio M-H, Melnichenko O, Centurioni L, Chambers D, Zlotnicki V, Galperin B (2009) Mean dynamic topography of the ocean derived from satellite and drifting buoy data using three different techniques. *J Atmos Ocean Technol*. doi:10.1175/2009JTECHO672.1, <http://www.springerlink.com/content/r882426635467007/>
58. Mayer-Gürr T, Eicker A, Kurtenbach E, Ilk K-H (2010) ITG-GRACE: global static and temporal gravity field models from GRACE data. *Adv Technol Earth Sci* 2010(Part 2):159–168. doi:10.1007/978-3-642-10228-8_13
59. Morison J, Wahr J, Kwok R, Peralta-Ferriz C (2007) Recent trends in Arctic Ocean mass distribution revealed by GRACE. *Geophys Res Lett* 34:L07602. doi:10.1029/2006GL029016
60. Muller PM, Sjogren WL (1968) Mascons: lunar mass concentrations. *Science* 161:680–684
61. Nerem RS, Wahr J (2011) Recent changes in the Earth's oblateness driven by Greenland and Antarctic ice mass loss. *Geophys Res Lett* 38:L13501. doi:10.1029/2011GL047879
62. O'Keefe JA, Eckels A, Squires RK (1959) The gravitational field of the earth. *Astron J* 64:245
63. Padman L, Howard SL, Orsi AH, Muench RD (2009) Tides of the northwestern Ross Sea and their impact on dense outflows of Antarctic Bottom Water. *Deep-Sea Res Pt II Oceanography* 56:818–834. <http://dx.doi.org/10.1016/j.dsr2.2008.10.026>
64. Park J, Watts DR, Donohue K, Jayne S (2008) A comparison of in situ bottom pressure array measurements with GRACE estimates in the Kuroshio extension. *Geophys Res Lett* 35:L17601. doi:10.1029/2008GL034778
65. Paulson A, Zhong S, Wahr J (2007) Inference of mantle viscosity from GRACE and relative sea level data. *Geophys J Internat* 171:497–508
66. Parker RL (1975) Theory of ideal bodies for gravity interpretation. *Geophys J Roy Astron Soc* 42(2):315–334
67. Peltier WR (2004) Global glacial isostasy and the surface of the ice-age Earth: the ICE-5G (VM2) model and GRACE. *Annu*

- Rev Earth Planet Sci 32:111–149. doi:10.1146/annurev.earth.32.082503.144359
68. Peltier WR, Luthcke SB (2009) On the origins of earth rotation anomalies: New insights on the basis of both “paleogeodetic” data and gravity recovery and climate experiment (GRACE) data. *J Geophys Res* 114:B11405. doi:10.1029/2009JB006352
 69. Peralta-Ferriz C, Morison J (2010) Understanding the annual cycle of the Arctic Ocean bottom pressure. *Geophys Res Lett* 37:L10603. doi:10.1029/2010GL042827
 70. Ponte RM, Quinn KJ (2009) Bottom pressure changes around Antarctica and wind-driven meridional flows. *Geophys Res Lett* 36:L13604. doi:10.1029/2009GL039060
 71. Quinn KJ, Ponte RM (2008) Estimating weights for the use of time-dependent gravity recovery and climate experiment data in constraining ocean models. *J Geophys Res* 113:C12013. doi:10.1029/2008JC004903
 72. Ray RD, Luthcke SB, Boy J-P (2009) Qualitative comparisons of global ocean tide models by analysis of intersatellite ranging data. *J Geophys Res* 114:C09017. doi:10.1029/2009JC005362
 73. Rietbroek R, LeGrand P, Wouters B, Lemoine J-M, Ramillien G, Hughes CW (2006) Comparison of in situ bottom pressure data with GRACE gravimetry in the Crozet-Kerguelen region. *Geophys Res Lett* 33:L21601. doi:10.1029/2006GL027452
 74. Rietbroek R, Brunnabend S-E, Dahle C, Kusche J, Flechtner F, Schröter J, Timmermann R (2009) Changes in total ocean mass derived from GRACE, GPS, and ocean modeling with weekly resolution. *J Geophys Res: Oceans* 114:C11004. doi:10.1029/2009JC005449
 75. Rignot E, Velicogna I, van den Broeke MR, Monaghan A, Lenaerts J (2011) Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise. *Geophys Res Lett* 38:L05503. doi:10.1029/2011GL046583
 76. Rio MH, Guinehut S, Larnicol G (2011) New CNES-CLS09 global mean dynamic topography computed from the combination of GRACE data, altimetry, and in situ measurements. *J Geophys Res* 116:C07018. doi:10.1029/2010JC006505
 77. Riva REM, Gunter BC, Urban TJ, Vermeersen BLA, Lindenbergh RC, Helsen MM, Bamber JL, van de Wal RSW, van den Broeke MR, Schutz BE (2009) Glacial isostatic adjustment over Antarctica from combined ICESat and GRACE satellite data. *Earth Planet Sci Lett* 288:516–523. <http://dx.doi.org/10.1016/j.epsl.2009.10.013>
 78. Rodell M, Velicogna I, Famiglietti JS (2009) Satellite-based estimates of groundwater depletion in India. *Nature* 460:999–1002. <http://dx.doi.org/10.1038/nature08238>
 79. Rowlands DD, Luthcke SB, McCarthy JJ, Klosko SM, Chinn DS, Lemoine FG, Boy J-P, Sabaka TJ (2010) Global mass flux solutions from GRACE: a comparison of parameter estimation strategies—mass concentrations versus stokes coefficients. *J Geophys Res* 115:B01403. doi:10.1029/2009JB006546
 80. Sasgen I, Martinec Z, Bamber J (2010) Combined GRACE and InSAR estimate of West Antarctic ice mass loss. *J Geophys Res (Earth Surface)* 115:F04010. doi:10.1029/2009JF001525
 81. Siegismund F, Romanova V, Köhl A, Stammer D (2011) Ocean bottom pressure variations estimated from gravity, nonsteric sea surface height and hydrodynamic model simulations. *J Geophys Res* 116:C07021. doi:10.1029/2010JC006727
 82. Sun AY, Green R, Rodell M, Swenson S (2010) Inferring aquifer storage parameters using satellite and in situ measurements: estimation under uncertainty. *Geophys Res Lett* 37:L10401. <http://dx.doi.org/10.1029/2010GL043231>
 83. Swenson S (2010) Assessing high-latitude winter precipitation from global precipitation analyses using GRACE. *J Hydrometeorol* 11:405–420. doi:10.1175/2009JHM1194.1
 84. Swenson S, Wahr J (2002) Methods for inferring regional surface-mass anomalies from gravity recovery and climate experiment (GRACE) measurements of time-variable gravity. *J Geophys Res* 107(B9):2193. doi:10.1029/2001JB000576
 85. Swenson S, Wahr J (2006) Post-processing removal of correlated errors in GRACE data. *Geophys Res Lett* 33:L08402. doi:10.1029/2005GL025285
 86. Swenson S, Wahr J (2009) Monitoring the water balance of Lake Victoria, East Africa, from space. *J Hydrol* 370(1–4):163–176. doi:10.1016/j.jhydrol.2009.03.008
 87. Swenson S, Chambers DP, Wahr J (2008) Estimating geocenter variations from a combination of GRACE and ocean model output. *J Geophys Res* 113. doi:10.1029/2007JB005338
 88. Tamisiea ME, Hill EM, Ponte RM, Davis JL, Velicogna I, Vinogradova NT (2010) Impact of self-attraction and loading on the annual cycle in sea level. *J Geophys Res* 115:C07004. doi:10.1029/2009JC005687
 89. Tapley BD et al (2004) GRACE measurements of mass variability in the Earth system. *Science* 305:503–505. doi:10.1126/science.1099192
 90. Tiwari VM, Wahr J, Swenson S (2009) Dwindling groundwater resources in Northern India, from satellite gravity observations. *Geophys Res Lett* 36:L18401. doi:10.1029/2009GL039401
 91. Tregoning P, Ramillien G, McQueen H, Zwartz D (2009) Glacial isostatic adjustment and nonstationary signals observed by GRACE. *J Geophys Res* 114:B06406. doi:10.1029/2008JB006161
 92. van den Broeke M, Bamber J, Ettema J, Rignot E, Schrama E, van de Berg WJ, van Meijgaard E, Velicogna I, Wouters B (2009) Partitioning recent Greenland mass loss. *Science* 326:984. doi:10.1126/science.1178176
 93. Velicogna I (2009) Increasing rates of ice mass loss from the Greenland and Antarctic ice sheets revealed by GRACE. *Geophys Res Lett* 36:L19503. doi:10.1029/2009GL040222
 94. Vianna ML, Menezes VV (2011) Double-celled subtropical gyre in the South Atlantic Ocean: means, trends, and interannual changes. *J Geophys Res* 116:C03024. doi:10.1029/2010JC006574
 95. Vinogradova NT, Ponte RM, Tamisiea ME, Quinn KJ, Hill EM, Davis JL (2011) Self-attraction and loading effects on ocean mass redistribution at monthly and longer time scales. *J Geophys Res-Oceans* 116. <http://dx.doi.org/10.1029/2011JC007037>
 96. Wahr J, Molenaar M, Bryan F (1998) Time variability of the Earth's gravity field: hydrological and oceanic effects and their possible detection using GRACE. *J Geophys Res* 103:30205–30229. doi:10.1029/98JB02844

97. Wahr J, Swenson S, Velicogna I (2006) Accuracy of GRACE mass estimates. *Geophys Res Lett* 33:L06401. doi:10.1029/2005GL025305
98. Willis JK, Lyman JW, Johnson GC, Gilson J (2008) In situ data biases and recent ocean heat content variability. *J Oceanic Atmosph Technol* 26:846–852. doi:10.1175/2008JTECHO608.1
99. Wouters B, Chambers DP (2010) Analysis of seasonal ocean bottom pressure variability in the Gulf of Thailand from GRACE. *Glob Planet Chang* 74:76–81. doi:10.1016/j.gloplacha.2010.08.002
100. Wu X, Heflin MB, Schotman H, Vermeersen BLA, Dong D, Gross RS, Ivins ER, Moore AW, Owen SE (2010) Simultaneous estimation of global present-day water transport and glacial isostatic adjustment. *Nat Geosci* 3:642–646. doi:10.1038/NNGEO938
101. Zaitchik BF, Rodell M, Reichle RH (2008) Assimilation of GRACE terrestrial water storage data into a land surface model: results for the Mississippi river basin. *J Hydrometeorol* 9:535–548. doi:10.1175/2007JHM951.1
102. Zlotnicki V, Wahr J, Fukumori I, Song Y-T (2007) Antarctic circumpolar current transport variability during 2003–2005 from GRACE. *J Physical Oceanog* 37(2):230–244

Books and Reviews

- Cazenave A, Chen J-L (2010) Time-variable gravity from space and present-day mass redistribution in the Earth system. *Earth Planet Sci Lett* 298(2010):263–274. doi:10.1016/j.epsl.2010.07.035
- Chambers DP, Schröter J (2011) Measuring ocean mass variability from satellite gravimetry. *J Geodyn* 52:333–343. doi:10.1016/j.jog.2011.04.004
- Dickey J et al (1997) *Satellite gravity and the geosphere*. US National Research Council, National Academy Press, Washington, DC

Green Catalytic Transformations

JAMES H. CLARK, JAMES W. COMERFORD, D. J. MACQUARRIE
Department of Chemistry, University of York,
Heslington, York, UK

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Zeolites

Clays
Supported Reagents
Future Directions
Bibliography

Glossary

Atom economy Calculates the percentage of atoms in the reagents used in the final product.

Chemisorption Where a species is chemically bound to a surface.

Clay A pure material, either synthesized or natural, used as a heterogeneous catalyst or support.

Dehydroxylation Where two isolated silanols condense to form a siloxane bridge, with loss of water.

E-factor A measure of waste produced per quantity of product.

EnvirocatsTM A series of commercially available catalysts developed by the University of York and Contract Chemicals Ltd.

Heterogeneous Where two species in a reaction, catalyst and reagents, are in a different phase, i.e., solid and vapor phase.

Homogeneous Where species in a reaction are in the same phase.

Ion exchange A process where ions, typically metal cations, are exchanged with other ions on a surface using a suitable solvent.

Physisorption Where a species interacts strongly with a surface through electrostatic interaction, hydrogen bonding.

Support/supported Referring to a species interacting with or bound to a surface.

Zeolite A diverse aluminosilicate structure commonly used in heterogeneous catalysis.

Definition of the Subject and Its Importance

With ever-increasing demand of chemical products on a global scale, as well as poor public image in recent years, there has been increasing pressure for chemistry industry to become more efficient and sustainable. Processes and catalytic cycles on large scales are being scrutinized by companies to enhance efficiency, reduce environmental impact and associated costs using state-of-the-art research and technology. An overwhelming number of new and improved catalytic

transformations are reported in many different fields on a daily basis. However, a catalytic transformation must fulfill a number of criteria to be deemed as “green.” The process/catalytic cycle must exhibit a notable improvement on existing syntheses, not only in terms of activity, but as an overall process by assessing waste produced throughout (cradle to grave concept including even the synthesis of the catalyst itself), potential reusability of catalyst, as well as ease of product isolation and potential user risk, where the safety/toxicity of chemicals involved in the process are assessed.

Introduction

The necessitated development of green synthetic procedures has grown from a number of different socio-economic factors over recent decades. Increasing demand of chemical products worldwide has meant that the environmental impact of industry (particularly within the EU), is being assessed more harshly than ever. This has had a distinct effect on the relationship between chemical industry and the environment, and as such green chemistry has received huge attention. The key factor of concern is the quantity and nature of the waste being produced, a number of different strategies have been implemented by the EU and ECHA to reduce waste by restricting the use of hazardous/potentially toxic chemicals on an industrial scale, through either banning the substance completely or imposing substantial tax deterrents on certain chemical wastes (REACH legislation). This provides a financial incentive for companies to employ waste-minimizing techniques for their processes. An excellent overview of waste minimization is given in “Chemistry of Waste Minimization” which gives a detailed insight into how environmental factors are affected by the production and restriction of chemical waste [1]. Focus has been increasingly diverted toward the prevention of waste rather than the treatment, through a combination of replacing homogeneous catalysts with heterogeneous catalysis, utilizing renewable raw materials, and investing in new technologies such as continuous flow.

Homogeneous Versus Heterogeneous

By definition, a heterogeneous catalyst is a catalytically active species that is in a different phase to the reagents

within a reaction system, more often than not a solid in either liquid or vapor phase synthesis. There are a number of advantages of using heterogeneous catalysts as opposed to the homogeneous equivalents, detailed below:

- **Safety** – An important consideration in the practice of green chemistry. Heterogeneous catalysts often tend to be environmentally benign and easy/safe to handle. This is due to the active species being adhered to a support (often forming a powder), essentially reducing its reactivity with the surrounding environment.
- **Reusability** – Due to the difference in phases, the catalyst is simply filtered off (through centrifugation on industrial scales) and reactivated for reuse many times over. For instance, zeolites used for petroleum refining can be reactivated and reused for many years – sometimes up to a decade – before disposal.
- **Activity** – In many cases increased activity is observed when supporting an active homogeneous species on a support, due to the complex but unique surface characteristics found with a variety of different supports.
- **Selectivity** – Heterogeneous catalysts can give an increased degree of selectivity in reaction pathway. This can simply be a consequence of adsorption of substrates and the consequential restricted freedom of movement of the reacting molecules. More famously, the pores of a solid support can cause size and shape restrictions with regard to the substrates, intermediates, and products. For example, the substitution of aromatic rings leads to products with different geometries and bulkier intermediates may be less likely to form and/or bulkier products may not be able to leave the pores leading to orientational selectivity. Shape selectivity can also affect stereoselectivity through control over reaction pathways.

However, there are a handful of disadvantages. The quantity of solid catalyst required is often higher than that of the homogeneous equivalent, due to the lower concentration on the support surface (reusability makes this less of an issue though). Blocking of the pores/support channels can occur with narrow pores sizes and reduce efficiency over time in some liquid phase reactions; nevertheless, this is often twinned with

high stereoselectivity. Despite this, the advantages found with heterogeneous catalysis far outweigh the minor drawbacks.

Green Chemistry Metrics

Green chemistry metrics are important tools in assessing the effectiveness and efficiency of reactions. Increasingly, they are becoming common place in the designing of synthetic routes/chemical syntheses, the main metrics are defined below:

- E-Factor – Developed by Sheldon, the environmental factor determines the quantity of waste (in Kg) produced per Kg of product. This takes into account all waste throughout the synthesis, not only solvents used in reactions, etc., but solvent used for recrystallisation. An E-Factor of 1 is considered ideal.

$$Ef = \frac{\text{Total Quantity of waste (Kg)}}{\text{Total quantity of product (Kg)}}$$

- Atom economy – Formulated by Trost, calculates the percentage of atoms in the starting material present in the final product, clean syntheses aim for 100%.

$$\text{Atom economy (\%)} = \frac{\text{Molecular weight of product}}{\text{Molecular weight of reagents}} \times 100$$

- Reaction mass efficiency – Developed by GSK, similar to atom economy but assesses difference in mass.

$$\text{Mass efficiency (\%)} = \frac{\text{Mass product}}{\text{Total mass reagents}} \times 100$$

- Carbon efficiency – Developed by GSK, again similar in format, determines difference in carbon mass between reagents and products.

$$\text{Carbon efficiency (\%)} = \frac{\text{Quantity of carbon in product}}{\text{Quantity of carbon in reagents}} \times 100$$

Overall, recent research has aimed to reduce the quantity and toxicity of waste produced by industrial processes and substantial focus has been centered on the type of catalysts used. Below is a detailed discussion of heterogeneous catalysts covering structural and mechanistic aspects as well as highlighting applications, interesting developments, and increased green credentials of many “classic” syntheses.

Zeolites

Background

The term “zeolite” was first employed by the Swedish mineralogist Cronstedt in 1756, literally meaning boiling stone in Greek [2]. Natural zeolites have had applications in chemistry over past centuries, such as odor control, gas separation, desiccants, water treatment, agriculture, etc., but the development of synthetic zeolites has been hallmarked as one of the greatest discoveries in science, allowing microporous structures to be “tailored to fit” a huge number of various applications such as extraction, purification, and heterogeneous catalysis. The replication of a natural zeolite was first achieved by Barrer in the early 1940s and further research of entirely synthetic zeolites was continued by Milton at the Union Carbide Linde Division’s Research Laboratory in Buffalo, New York, in 1949. The developed Zeolite A and B(P) were used for the separation of gases based on size, in particular air, to produce high purity oxygen for steel mills and other large scale applications [3], although this was not commercialized until the 1970s. Following research by Mobil focused on zeolites for large scale petroleum cracking and dozens of zeolites were produced that were able to act as acid catalysts.

Composition and Structure Zeolites are crystalline three dimensional microporous structures with tetrahedral building blocks comprised of $[\text{SiO}_4]^{-4}$ and $[\text{AlO}_4]^{-5}$. The chemical composition is often represented by $\text{M}_{2/n} \text{O} \cdot \text{Al}_2\text{O}_3 \cdot y\text{SiO}_2 \cdot w\text{H}_2\text{O}$, where, due to the trivalent aluminum species, the zeolite has a negative charge proportional to each Al and requires either a pentavalent element, such as a neighboring phosphorus (P^{5+}) or a cation to stabilize this. Non-framework cations are found in the porous network and can vary; alkaline metals from group IA and IIA such as Na^+ are often used. The tetrahedral building blocks form structures known as secondary building units, (SBU), of which there is a large variety (around 16). It is these units that form the larger unit cells (cages) which make up the zeolitic framework, commonly used to classify zeolites, although they can be classified through the type of SBU, framework density (the number of T atoms per unit cell), or pore size. For instance, sodalite (SOD), faujasite

(FAU), and zeolite A (LTA) share the same sodalite unit cell, however the SOD cages are of a cubic arrangement, LTA is similar in arrangement but linked together through the 4–4 SBU and FAU comprises of a diamond like structure of sodalite cages, linked together through the hexagonal face [4], Fig. 1. The specific naming of the frameworks is complex and will not be discussed.

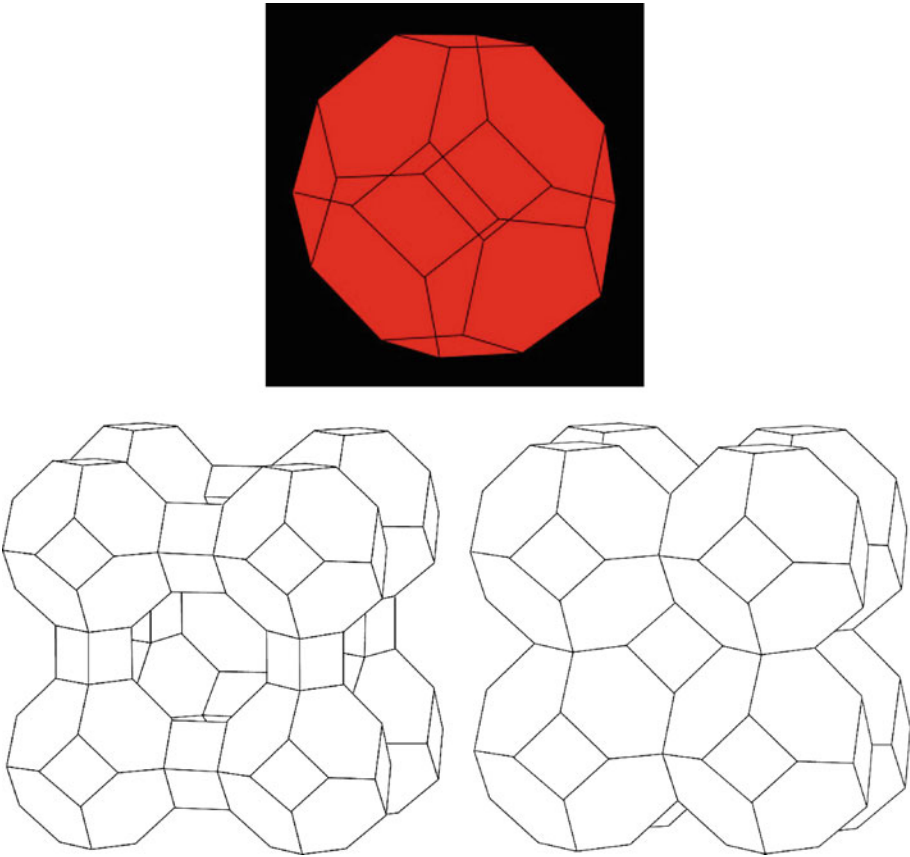
Another common classification mentioned above was that of aperture size, referring to the diameter of the rings forming the zeolite channel. Table 1 shows a range of zeolite spanning the typical pore sizes from small to ultralarge, found with common zeolites [1].

Pore sizes up to $\sim 14\text{\AA}$ are categorized as microporous in accordance with the IUPAC classification of materials [5]. The size and shape of the zeolite for catalytic applications is crucial, potentially giving

exceptional shape size selectivity. Aside from pairing framework with application, it is possible to tune the pore size by changing the Si:Al ratio; a reduction will produce a smaller unit cell, yet require fewer stabilizing

Green Catalytic Transformations. Table 1 Pore sizes of common zeolites

Zeolite	Number of tetrahedral in ring	Diameter of main channels \AA
Sodalite	4	very small – <4
Zeolite A	8	Small – 4
ZSM-5	10	Medium – 5.6
Faujasite	12	Large – 7.4
Cloverlite	20	Ultra large – 13.2
MCM-41	>20	Macroporous – 100



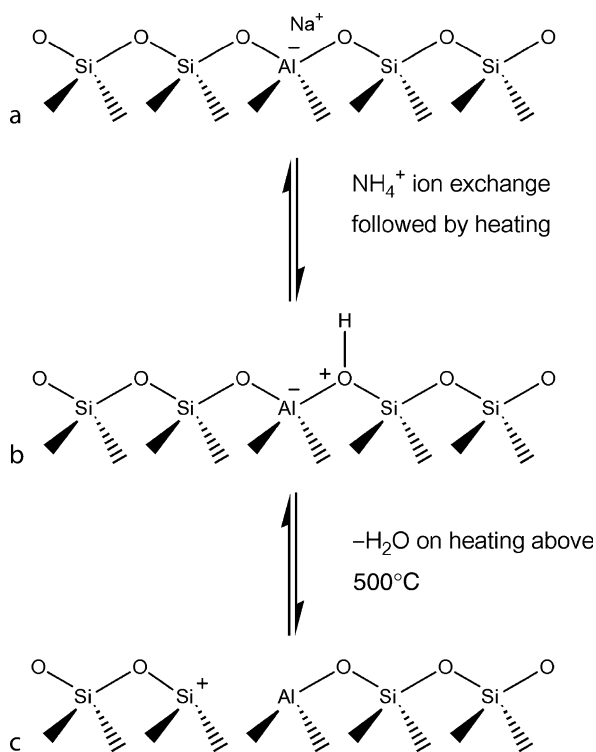
Green Catalytic Transformations. Figure 1
Top – SOD unit cell; *Left* – LTA framework; *Right* – SOD framework

cations, freeing up the zeolite channel. In addition, modification post-synthesis is possible simply through ion exchange, the choice of ions can increase the effective size of the pore opening. Despite the selectivity obtained with smaller pore size zeolites, there can be diffusional and blocking problems in liquid phase reactions. As a result there has been much research into mesoporous zeolites, allowing increased diffusional rates and reduced deactivation.

Synthesis The synthesis of zeolites typically involves mixing a silica and alumina source under basic conditions. Also incorporated into the mixture is a suitable cation of choice, such as either an alkaline metal in an oxide, hydroxide, or salt, with dual function either as counterions in the final material or acting as structure directing templates. Aside from the properties of the mixture, temperature, pressure, and crystallization time are also important variables and vary considerably. Typical syntheses crystallize between room temperature to 200°C under autoclave pressure, the time required ranging from hours to days.

Acidity Counterions within the porous network (Fig. 2a) can be readily exchanged using salt solutions, a process referred to as *ion exchange*. To give the zeolite Bronsted acidity, ammonium salt solution (such as ammonium hydroxide) can be used to exchange the metal ions with ammonium. On heating the material deprotonates giving hydrogen as the stabilizing cation (Fig. 2b). The Bronsted acid strength is dependent on the Si:Al ratio, where a high ratio of silicon to aluminum results in stronger acidity, as well as increased hydrophobicity, i.e., zeolite B150 is a stronger solid acid and more hydrophobic than B25 (with a ratio of Si 25: Al 1). Further heating above 500°C results in the reversible dehydroxylation of the silanol adjacent to the Al site, giving an aluminum Lewis acid site and a charge silicon species (Fig. 2c).

The basicity of zeolite tends to be less well documented. Lewis basicity stems from a negative charge on the oxygen where Bronsted is due to an extraframework cation on the surface [3].



Green Catalytic Transformations. Figure 2

Ion exchanged zeolite and subsequent heating to give Lewis basicity. (a) Zeolite as synthesized; (b) Bronsted acid zeolite; (c) Lewis acid zeolite

Catalysis

There are an overwhelming number of examples of solid zeolite acid, base, and oxidation catalysis throughout the past 60 years. As such, the following section will discuss the most popular applications throughout the decades as well as recently published examples. A common observation with much of the research is the impressive shape selectivity and structural diversity found with zeolitic catalysis.

Catalytic Cracking It has been around 50 years since Mobil first developed the fluid catalytic cracking (FCC) process using zeolite technology and this still remains the dominant catalytic application for zeolites today. Nevertheless, there are many other important transformations using zeolites, of which, a selection are

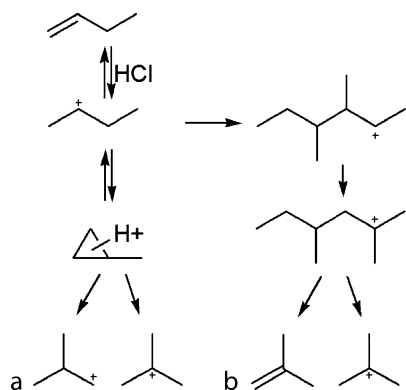
Green Catalytic Transformations. Table 2 Showing applications and corresponding zeolites

Application	Zeolite	Notes
Catalytic Cracking	HY-high silica zeolite REY Medium pore zeolites	Selectivity and high conversion rate
Hydrocracking	H-ZSM-5, Faujasite (X, Y), metals on Mordenite (Co, Mo, W, Ni) USY, CAMgY	Conversion rate high
Dewaxing	Pt-mordenite, ZSM-5, silicalite, high silica zeolites (ferrierite)	Low pour and cloud points
Hydroisomerisation	Pt-Mordenite	Low octane pentanes and hexanes converted to high octane yields
Aromatization	Pt, K, Ba silicalite, Pt, Ga, Zn-ZSM-5	Aromatization of C3–C8 cuts, aromatization of LPG
Benzene alkylation	ZSM-5	Production of ethylbenzene and styrene with low quantity of by-products
Xylene isomerisation	ZSM-5	High selectivity to para yield
Toluene disproportionation	ZSM-5, Mordenite	Production of xylenes and benzene
Methanol to petroleum	ZSM-5, erionite	High olefin yields (high octane rating) and high petroleum yield
Fischer-Tropsch	Metal-ZSM-5 (Co, Fe)	Natural gas to petroleum

shown in Table 2 [3]. Catalytic cracking typically employs fluidized catalyst bed technology and is simple in concept, but a detailed insight of the process and chemistry is outside the scope of this encyclopedia. As an example take the basics of petroleum refining; the initial feedstock comprises a range of carbons chains/aromatics (paraffins, naphthenic and aromatic hydrocarbons), light being C₁₀–C₂₀ and heavy being C₁₅–C₂₅. This is pre-heated to ~370–400°C and mixed with powdered solid catalyst at a ratio of 2:1 oil:catalyst, to give a slurry which is injected into a reaction vessel at ~480–550°C [6]. At this temperature the oil partially evaporates and reacts with the catalyst, after 1–4 s the mixture is separated into catalyst, product for distillation and unreacted product (~20–25% of the starting material). The deactivated catalyst is simply reactivated in a stream of air ~700°C and re introduced back into the process. Impressively, the catalyst can be reused for many years in this highly efficient continuous process. The most utilized catalysts currently used in production consist of type Y faujasite zeolite with a Si:Al ratio of ~2.5 and ZSM-5 with a Si:Al ratio of ~15 to >100, allowing control of activity and selectivity. Typical stabilizing cations consist of H⁺, Na⁺, Ca²⁺, Ba²⁺, NH₄⁺, NR₄⁺ and various other rare earth metals. Despite some of their advantageous properties to the zeolite structure, Na⁺ and particularly K⁺ act as a catalyst poisons.

The ZSM-5 catalyst has since in developed in the field of FCC and cationic exchange with phosphorus has been of recent interest [7–13]. Enhanced activity was reported [14] in the cracking of propylene in the presence of phosphorus exchanged ZSM-5 prepared through hydrothermal dispersion. Classic modification techniques such as impregnation and ion exchange result in a material that can easily lose phosphorus, leading to contamination and loss of activity/selectivity. It is thought that the phosphorus compounds interact with bridged OH groups on the surface reducing the zeolites acidity, consequently increasing activity and modifying shape selectivity.

Isomerization Skeletal isomerisation again plays an important role in the refining industry. A well researched transformation is that of n-butene into



Green Catalytic Transformations. Figure 3

Isomerisation of butene to isobutene. (a) Monomolecular mechanism with cyclopropanic intermediate proposed by Brouwer and Hogeveen [15]; (b) Bimolecular mechanism proposed by Guisnet et al. [16]

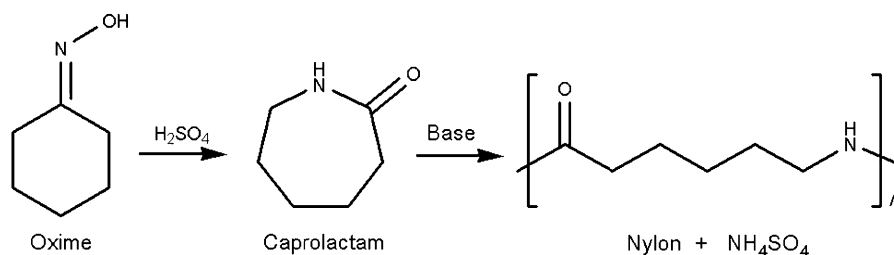
isobutene, an intermediate to methyl-tert butyl ether (MTBE), one of the most important oxygenated additives used in lead free petroleum, Fig. 3. Initially the reaction is thought to be acid catalyzed consisting of a double bond shift to 2-butene, where the second isomerisation step to isobutene requires a stronger acid. Conversely, an increase in acid strength, as found with classic methods employing homogeneous acids such as HCl, also increases quantity of by-products, such as dimerization to octenes, cracking to C₂, C₃, C₅ and C₆ olefins and formation of coke [17]. Potential catalysts, therefore, were of medium acidity to increase selectivity to isobutene and reduced deactivation. As such there has been much research into ferrierite (FER), which shown extremely high selectivity under mild conditions [18–22].

Direct comparison between FER and zeolites of similar acidity (i.e., ZSM-5 – MFI framework) has shown the former to be consistently more selective. This has been ascribed to two main factors; firstly the low number of acid sites (high Si/Al ratio) is thought to prevent undesired bimolecular processes [6] such as aromatization, cyclization, hydrogen transfer and oligomerization, reducing by-products and coke formation (therefore increasing selectivity) and secondly the unique structure of ferrierite. However, despite the high performance of this catalyst, the role of coke formation as well as the precise mechanism in the isomerisation of butane to isobutene is still in debate [19].

Rearrangement The synthesis of ϵ -caprolactam through the Beckmann rearrangement of cyclohexanone oxime is industrially important in the synthesis of nylon-6 and resins, Fig. 4. Previously, the synthesis would involve the conversion of cyclohexanone with hydroxylamine sulfate or phosphate, to give cyclohexanone oxime. Subsequently, strong acids such as fuming sulphuric were used as an acid catalyst, followed by treatment with ammonia yielding the final ϵ -caprolactam product. However, large quantities of (NH₄)₂SO₄ salt by-product are formed, along with a high reactor corrosion factor and high risk factor (handling large volumes of concentrated acid), making the process environmentally unsound.

As such this has been studied intensely, with a number of developed catalysts being used to varying degrees of success, Table 3. Out these, titanium silicate (TS-1) has shown exceptional activity and selectivity, its development over the past decade gives an excellent example of green synthesis. The old two step process of ammoximation followed by rearrangement has been telescoped into a single continuous flow process [28] and has been used industrially by Sumitomo Chemical Co., Ltd since 2003, (operating on 60,000 tons per year scale!). TS-1 is used to convert cyclohexanone [29] to the oxime giving a conversion of >99% with a selectivity of 98%. This is followed by vapor phase rearrangement of the oxime to the lactam using high silica MFI zeolite, giving a yield on >99% with a selectivity of 95%. The improvement of this reaction can be shown by the increase in atom economy for ammoximation, from 36% using H₂SO₄ + 1.5NH₃ to 100% using TS-1 [11].

Schüth recently reported [30] that by crosslinking the TS-1 structure, both improved catalytic activity and reduced deactivation over time was observed. The preparation of the zeolite uses siloxane linkers, resulting in an increased in mesoporous surface character similar to that of silicalite. Interestingly activity correlates with increased concentration of linkers, as well as Ti concentration. Deactivation of that catalyst was found to be substantially improved compared with standard TS-1, up to 32% increased productivity. The observed effects are thought to be due to both the presence of titanium and siloxane crosslinkage, cooperatively reducing coke formation. The benefits of these heterogeneous catalysts aside from high activity and selectivity include simple

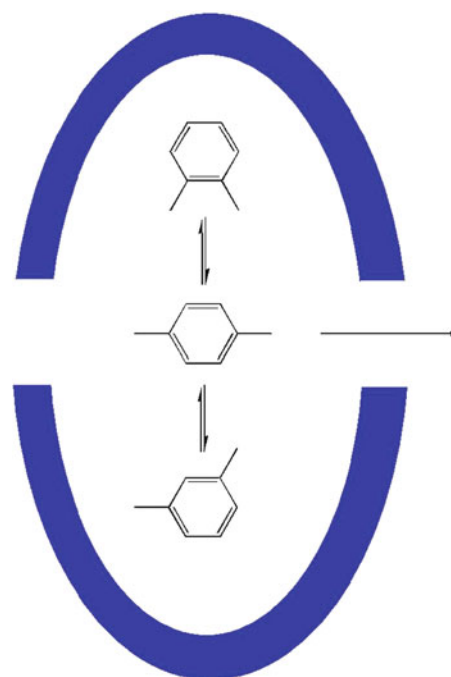
**Green Catalytic Transformations. Figure 4**Rearrangement of oxime to give ϵ -caprolactam and nylon**Green Catalytic Transformations. Table 3** Examples of catalysts reported for the synthesis of ϵ -caprolactam [23–27]

Catalyst	TOS ^a	Yield (%)	Selectivity (%)
FSM-16	0.5 h	98	49
ZnO/FSM-16	"	99	69
Al ₂ O ₃ /FSM-16	"	98	63
Si-MCM-41	3 h	>99	32
Al-MCM-41(A) ^b	"	>99	48
Al-MCM-41(B)	"	>99	65
Al-MCM-41(C)	"	>99	87
β -MFI	4 h	95	95
H ₃ BO ₃ /H β	"	>99	80
H-LTL	6 h	>99	97
H-OFF-ERI	"	>99	96
Silicalite-1 MFI	"	77	95
High Si MFI	"	>99	95
H-USY	"	>99	82
H-MOR	"	92	90

^aTime on stream^bIncreasing Si:Al ratio

separation of product and any by-products though filtration, recovery, reuse, and low toxicity.

Transformations of Aromatics over Zeolites The production of ethylbenzene from benzene and ethylene is an important step in the synthesis of styrene and subsequently polystyrene [3]. Previously the process used AlCl₃ in a Friedel-Crafts alkylation and as with many strong homogeneous Lewis acids, there are

**Green Catalytic Transformations. Figure 5**
Restrictive ZSM-5 channels

problems with corrosion, waste separation, and contamination. A cleaner, continuous process using fixed beds of ZSM-5 gives high yields. In particular, ZSM-5 is an excellent example of shape selective properties and tuneability available with zeolite catalysis. The alkylation of toluene and methanol, for example, gives p-xylene, industrially important for its use in the manufacture of terephthalic acid and dimethyl terephthalate (starting material for polyester fibers, vitamins and other pharmaceuticals) [1]. The exceptional selectivity observed with ZSM-5 is due to the different diffusional rates of the ortho, meta, and para isomers through the porous zeolite channels, Fig. 5. It has been

demonstrated through variable temperature reactions that the diffusion of p-xylene is around 1,000 times faster than the ortho and meta isomers [31].

Despite this high selectivity, further research into ion exchange of ZSM-5 has produced even more active catalysts. By exchanging ions of different sizes into the porous network the size of the channels can vary, becoming increasingly restrictive with large metal ions. Sotelo [32] discusses ion exchange with magnesium in H-ZSM-5, where para selectivity approaches 100%, with isomer by-products being due to reaction on the external surface of the catalyst.

Zeolites in Fine Synthesis of Speciality Chemicals

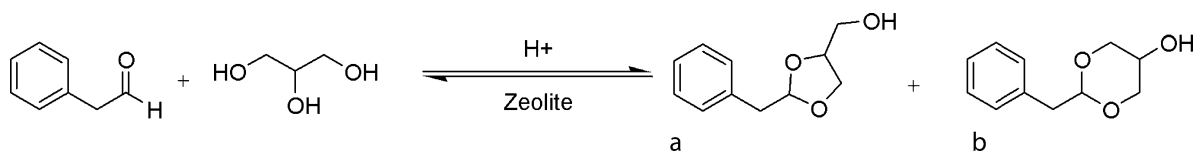
As well as their application to a number cracking, fuel and polymer reactions, zeolites are also able to catalyze a number of syntheses in fine chemical industry, such as in the synthesis of perfumes, fragrances, flavors, and pharmaceuticals, a few examples are given below.

Fragrances Vanillin propylene glycol acetal and phenylacetaldehyde glycerol acetals are important flavoring compounds with a vanilla and hyacinth fragrance. The propylene glycol acetal of vanillin is often used to imitate vanilla flavors as it causes flavor attenuation, this can easily be reversed through hydration back to the aldehyde, restoring full flavor [33]. Typical homogeneous acid catalysts are often used in the common transformation of the aldehyde to the acetal and suffer similar drawbacks with waste, separation, necessitated neutralization, and expense. A number of zeolites have therefore been applied to many fragrant syntheses, such as Jasminaldehyde (*n*-amylcinnamaldehyde), a violet scent obtained through the condensation of heptanal and benzaldehyde [34] and Fructone (ethyl 3,3-ethylendioxymethylbutyrate), a strong fruity scent of apple/pineapple/wood synthesized thorough the acetalization of ethyl acetoacetate with ethylene

glycol [35]. Climent et al. reported of synthesizing 2-benzyl-4-hydroxymethyl-1,3-dioxolane (Fig. 6a) and 2-benzyl-5-hydroxy-1,3-dioxane (Fig. 6b) from the acetalization of glycerol (soon to be a by-product on a million ton/pa scale) and benzaldehyde [33], both give a hyacinth scent.

Out of a number of zeolites investigated (Table 4), USY-2 and beta were particularly active and selective, comparable to p-toluene sulphonic acid. The secondary product (B) is thought to be converted from the acid catalyzed rearrangement of the less stable (A). Interestingly, the ratio of product to by-product is lower for Mordenite and ZSM-5, though overall conversion is low, again demonstrating shape selectivity and narrow pore size of the zeolites. Selectivity is not of huge importance in this case as both are active fragrances and are industrially acceptable, despite A having slightly increased potency than B.

Pharmaceuticals Ibuprofen, [(±)-2-(4-isobutylphenyl)] propionic acid, is a readily available and commonly used nonsteroidal anti-inflammatory/pain killer, able to treat many ailments such as muscular injury, headaches, and cold and flu symptoms (the list goes on). Its synthesis comprises of a six step process first patented by Boots in the 1960s. Typically isobutylbenzene is acylated with acetic anhydride using AlCl_3 as a Lewis acid catalyst; this causes problems with waste and to a certain extent selectivity, only isobutyl acting as a directing group to the para position (the ortho and meta are observed less so due to steric and electron effects, respectively). The resulting 2-(4-isobutyl phenyl)-ethanone is then reacted with Propionic acid (1-chloro-ethyl ester) and sodium ethoxide resulting in an epoxide, which, under acid catalysis, eliminates methyl ethanoate to give 2-(4-Isobutyl-phenyl)-propionaldehyde. There has been research into the replacement of AlCl_3 with



Green Catalytic Transformations. Figure 6

Synthesis of 2-benzyl-4-hydroxymethyl-1,3-dioxolane and 2-benzyl-5-hydroxy-1,3-dioxane

zeolites for synthesis of the intermediate 2-(4-isobutyl phenyl)-ethanone. Beta zeolite is active as an acylation catalyst and interestingly increased surface area through grinding or from formation of small particle sizes gives increased yields to around 20–30% [36, 37]. Higher yields, however, are observed in the synthesis of 2-acetyl-6-methoxynaphthalene ethanone, a precursor to (2-(6-methoxy-2-naphthyl)propionic acid) otherwise known as Naproxen, a similar nonsteroidal anti-inflammatory [38, 39].

Despite the fact that the 1-acetyl-2-methoxynaphthalene (A) Fig. 7 is kinetically favored [40], beta zeolite channels show selectivity toward the 2-acetyl-6-methoxynaphthalene (B). The increased selectivity is generated by the narrow pores, subsequently giving lower yields, potentially due to blocking, Table 5 [36].

After modification (grinding) an increased yield is observed thought to be due to an increase in surface area and accessibility to the acidic sites previously deep within the porous network. However, without the restrictive nature of the pores, selectivity to the 2-acetyl-6-methoxynaphthalene decreases, suggesting that the

synthesis of the 1-acetyl-2-methoxynaphthalene largely occurs on the external surface of the zeolite. The incorporation of Ce^{3+} gives higher selectivity thought to be due to increased Lewis acidity.

Limitations of Zeolites as Catalysts Despite the numerous advantages found with zeolite catalysis there are still a handful of downsides. Most zeolites tend to have an aperture size between 4 and 8 Å which limits application to less bulky substituents, mesoporous zeolites alleviate this for the most part, yet they too have their own disadvantages. The formation of coke (known as coking) as well as the buildup of products/reagents can cause rapid deactivation of the zeolite. The cost of the zeolite is often higher than that of an equivalent homogeneous acid, but as the cost of disposal of toxic chemical waste increases this will become less of an issue.

Clays

Background

Clays are a class of soil <2 μm in diameter primarily composed of fine-grained phyllosilicates (Greek for leaf, “phylon” and Latin for flint, “silic”), which when wet is generally plastic and sticky in appearance and can be dried out to give a hard, cohesive material [41]. The large majority of applications in the catalytic field are focused on acid catalysis. A particularly helpful book is “The Origin of Clay Minerals in Soils and Weathered Rocks” with a detailed discussion of clay structure and type [42].

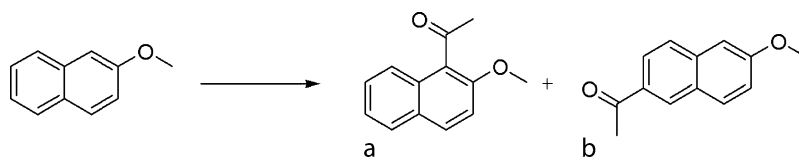
Structure of Clays

The subject of clay structure is a complex one [2]; as this chapter is designed to give an overview of clean synthesis, the following discussion will focus on the

Green Catalytic Transformations. Table 4 Conversion and selectivity to acetal product using various zeolites

Catalysts	Conversion (%) ^a	Selectivity (%)	
		A	B
USY-2	93	58	35
Beta-2	92	61	31
Mordenite	33	28	5
MCM-41	36	26	10
ZSM-5	54	46	8
PTSA	97	66	31

^aAfter 1 h in refluxing toluene



Green Catalytic Transformations. Figure 7

1-acetyl-2-methoxynaphthalene and the 2-acetyl-6-methoxynaphthalene precursors

Green Catalytic Transformations. Table 5 Acylation of isobutylbenzene with acetic anhydride

Catalyst	Yield (%)	Product distribution	
		1-Ac-2-Mn (A)	6-Ac-2-Mn (B)
Beta zeolite	68	30	70
Beta zeolite ^a	75	8	92
Beta zeolite ^b	75	24	76
BZ-1 ^c	81	35	65
BZ-1	86	25	75
BZ-2 ^d	82	35	65
Ce ³⁺ -BZ-2 ^e	88	22	78

^aUsing propionic anhydride

^bPre-activation at 750°C

^cMicrocrystalline beta zeolite of particle size 1–10 µm, obtained through mechanical disintegration

^dMicrocrystalline beta zeolite of particle size 10–50 µm obtained through shortening the crystallization time to 48 h instead of one week

^eIon exchanged zeolite using 5%wt metal chloride solution

basics. Clay structure consists of sheets, tetrahedral and octahedral, which together can form layers, 1:1 and 2:1. The tetrahedral sheet is made up of $[\text{SiO}_4]^{-4}$ groups linked together through three *basal* oxygens, the fourth however is termed apical and points away from the layer toward the adjacent octahedral layer. The basal “structural” oxygens form a hexagonal lattice, characteristic of the tetrahedral layer. It should be noted that through ionic substitution, the Si^{4+} center can be replaced with a cation of similar size, such as Fe^{3+} or Al^{3+} . The octahedral layer comprises of $\text{AlO}_4(\text{OH})_2$ units in sixfold coordination (although the cation center can be Al^{3+} , Fe^{3+} , Fe^{2+} , or Mg^{2+} depending on the clay). Depending on the oxidation state of the metal center (2+ or 3+), the structure can fall into a further two categories, dioctahedral or trioctahedral. The former type refers to the state where only two out of a possible three cation sites are filled, as a result of a divalent cation, this is often referred to as a gibbsite sheet (based on the naturally occurring material). Trioctahedral clays have trivalent cations; subsequently all potential cationic sites are filled.

The sheets can form layers in two ways (generally as there are exceptions), 1:1 and 2:1. The former refers to

a single tetrahedral sheet bonding with an octahedral, whereas 2:1 refers to an octahedral sheet “sandwiched” between two tetrahedral sheets (Fig. 8).

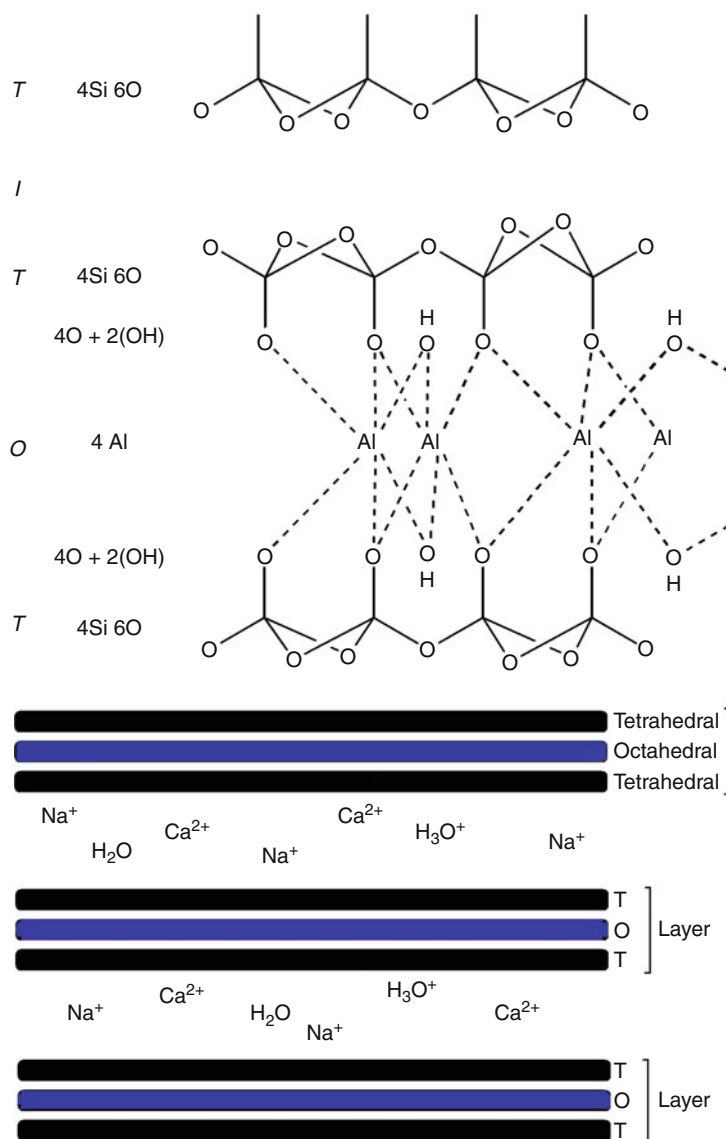
The sheets bond through the apical oxygen via the interlayer cations, hydrogen bonding, or van der Waals/electrostatic forces. Furthermore, the spatial arrangement of the tetrahedral and octahedral sheets is not directly compatible, creating distortions in either one or both structures (classed as the polytype); however, this level of detail is outside the scope of this encyclopedia. The region between the layers is known as the *interlayer*, (X). This provides access to sites within the clay structure and contains equivalents of stabilizing ions to counteract charge created by isomorphic substitution within the tetrahedral layer, key to the clays activity. Lastly, layers found in various clays can be mixed, in that they contain different octahedral layers, i.e., one layer of chlorite and one of smectite, in regular or irregular quantities. Table 6 shows the structural makeup of some 1:1 and 2:1 clays.

Montmorillonite

Out of the diverse range of clay types available, montmorillonite is the most utilized when applied to organic catalysis. It is comprised of a 2:1 structure of tetrahedral coordinated silicate $[\text{SiO}_4]^{-4}$ and octahedrally coordinated gibbsite $[\text{Al}_2(\text{OH})_6]$. The material is particularly susceptible to isomorphic exchange, where Al^{3+} replaces Si^{4+} in the tetrahedral sheet and Mg^{2+} replaces Al^{3+} in the octahedral sheet. This creates a negatively charged layer that is compensated by interchangeable cations, such as Na^+ and Ca^{2+} , which are situated between the layers. In fact, large quantities of cations can be “held” or “stored” in the clay depending on the extent of isomorphic exchange. On hydration the clay swells as the layers separate away from each other, allowing easy exchange of stabilizing cations in and out of the clay.

Acidic Properties of Clays

Clays can exhibit both Bronsted and Lewis acidity, the latter being due either to structural cations on the surface of the sheets or ions exchanged in the interlayer region. The Bronsted acidity stems from strong dissociation of intercalated water molecules coordinated to



Green Catalytic Transformations. Figure 8

Structure of a 2:1 clay

the Lewis acid center, generating mobile H-bonded protons in a highly polarizing environment, given as,



It follows that the weaker the Lewis acid, i.e., the more electron withdrawing the cation, the more acidic the Bronsted acid [45]. Despite being able to reach acidities close to that of 98% H₂SO₄, increased

acidity as well as surface area and hydroxyl group concentration have been researched resulting in a common process, the acid treatment of clays. Typical acid treatment involves using a strong inorganic acid such as HCl, sulphuric, or phosphoric in various quantities. This not only replaces exchangeable cations with hydrogen, but also leaches Al out of the central octahedral layer, producing enhanced surface area and increased acidity [46, 47].

Green Catalytic Transformations. Table 6 Composition of 1:2 and 2:1 clays [43, 44]

Group	Interlayer	Diocahedral	Triocahedral
1:1	None or only H ₂ O	Kaolinite Al ₂ Si ₂ (O ₅)(OH) ₄	Serpentine Mg ₃ Si ₂ (O ₅)(OH) ₄
2:1	None	Pyrophyllite Al ₂ Si ₄ (O ₁₀)(OH) ₂	Talc Mg ₃ Si ₄ (O ₁₀)(OH) ₂
Smectite 0.25 < x < 0.6	Hydrated exchange-able cations	Montmorillonite M _x [Al _{2-x} Mg _x](Si ₄)O ₁₀ (OH) ₂	Hectorite M _x [Mg _{3-x} Li _x](Si ₄)O ₁₀ (OH) ₂
		Beidellite M _x [Al ₂](Si _{4-x} Al _x)O ₁₀ (OH) ₂	Saponite M _x [Mg ₃](Si _{4-x} Al _x)O ₁₀ (OH) ₂
Vermiculite 0.6 < x < 0.9		Vermiculite (DIO) M _x [Al ₂](Si _{4-x} Al _x)O ₁₀ (OH) ₂	Vermiculite (TIO) M _x [Mg ₃](Si _{4-x} Al _x)O ₁₀ (OH) ₂
Mica	Non-hydrated cations	Muscovite K[Al ₂](Si ₃ Al)O ₁₀ (OH) ₂	Phlogopite K[Mg ₃](Si ₃ Al)O ₁₀ (OH) ₂

Pillared Clays

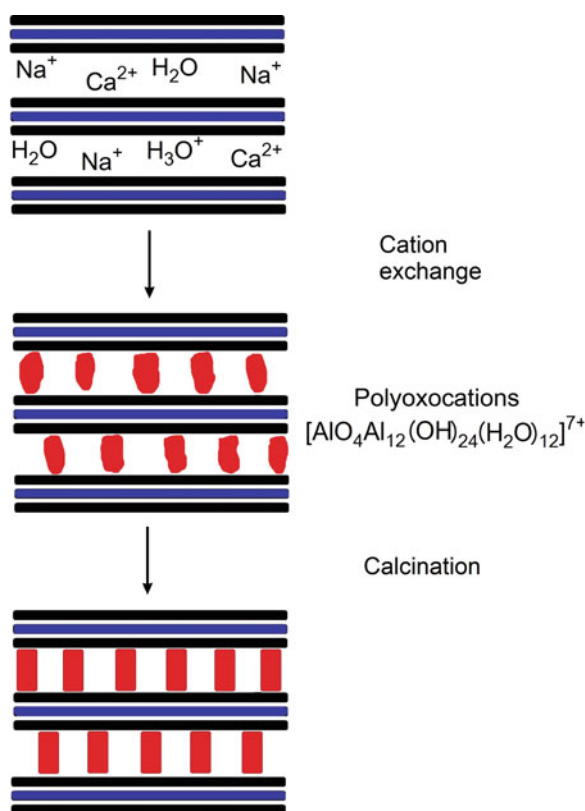
Despite high acid strength and good surface areas, basic and acid treated clays are still susceptible to thermal degradation; structural collapse of the clay is typically observed upon heating above ~200°C. By *pillaring* the clays increased surface area, concentration of acid site, and thermal stability is achieved, due to a fixed metal oxide pillar. A common pillaring agent is polyoxocation Al₁₃, prepared by mixing a base with aqueous chlorohydrate, AlCl₃, or Al(NO₃)₃ to give a solution with OH/Al³⁺ ratios of up to 2.5. The polyoxocation complex has been reported to have a tridecamer structure of [AlO₄Al₁₂(OH)₂₄(H₂O)₁₂]⁷⁺, also known as the Keggin ion [48]. The complex is then mixed with the clay, allowing diffusion of the structure between the layers and ion exchange. Once the clay is then filtered, washed, and calcined between 300°C and 500°C, the cations become bound supporting pillars and are able to release protons, Fig. 9.

Catalysis

As with zeolite, clays have found many applications, due to their high acidity, surface areas, high temperature tolerance (pillar only), and unique swelling properties. Discussed below are a few classic examples as well as more recent developments.

Esterification Organic esters are important in a plethora of applications, from intermediates,

pharmaceuticals to fragrances and perfumes to plastisizers (the list goes on). However, there are a number of downsides to conventional esterification process. If using a non-activated carboxylic acid, for instance, the reaction has to be catalyzed with a strong acid in the presence of a large quantity of alcohol, due to the reversible nature of the reaction. It is often the case where the acid is first activated, where electron density is withdrawn from the acid center, usually using thionyl chloride to produce an acid chloride. On reacting the alcohol with the acid chloride a large quantity of HCl waste is produced and a base is often added to the reaction mixture to “mop up” the inorganic acid. Overall the synthesis produces either a large quantity of alcohol or acid waste, or has more than a single step, activation of the acid with a toxic and potent lachrymator, thionyl chloride (PCl₅ has also been used). Avoid the use of these agents not only has a positive environmental impact, but reduces worker risk, and substantially reduces cost. Consequently there has been much research into alternative solid acids and Al³⁺ exchanged montmorillonite clay in particular has shown promising developments. Recent investigation [49, 50] into the synthesis of p-cresyl phenylacetates (industrially used for perfumes, soaps, etc.), using Al³⁺ exchange montmorillonites, has been reported, Table 7. The use of the p-cresyl, and aromatic alcohols in general, is of interest as they are weakly nucleophilic due to lone pair electron delocalisation around the ring, therefore requiring a strong delta positive charge on the acid center.



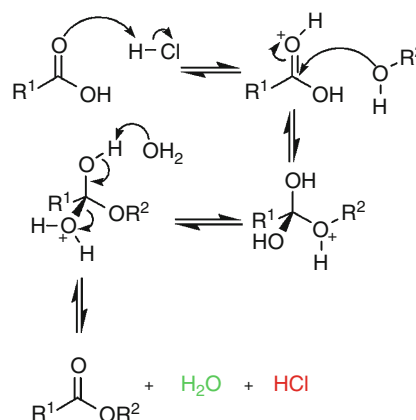
Green Catalytic Transformations. Figure 9
Pillared clays

Green Catalytic Transformations. Table 7 Synthesis of p-cresyl phenylacetate using various montmorillonite catalysts

Catalyst ^a	Calcination temperature (°C)	Reaction time (hours)	Yield (%)
Na ⁺	100	16	Nil
H ⁺	100	12	52
Al ³⁺	100	6	67
Al ³⁺	200	12	36
Al ³⁺	400	12	Nil

^aIon exchange on montmorillonite clay

Interestingly, the Na⁺ exchanged (raw) clay is not active whatsoever and is thought to be attributed to the low electronegativity of the Na⁺ ions being less able to polarize the interlamellar water and produce strong bronsted acidity. When exchanged with hydrogen, however, increased yields are observed. It is only with



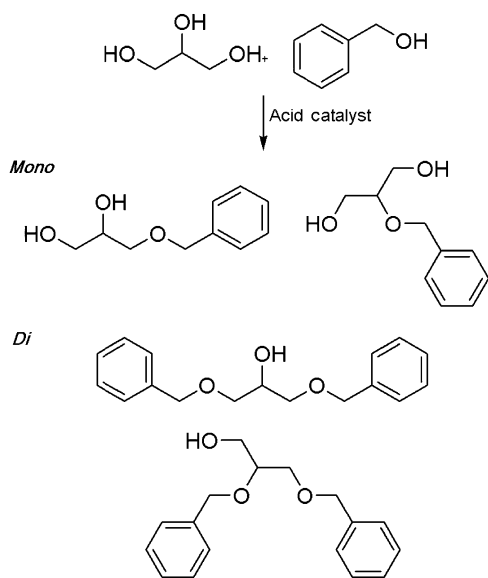
Green Catalytic Transformations. Figure 10
Mechanism of acid catalyzed esterification

the highly polarizing Al³⁺ cation exchange that the highest yields are seen, due to the high acid strength required to protonate and give the oxonium ion intermediate [10], Fig. 10. Increased calcination temperatures cause loss of water from the catalyst surface, converting Bronsted into Lewis acidity, resulting in a loss of activity.

This is an excellent example of a clean synthesis, due to its low molar quantity of alcohol, one part acid to two parts alcohol, the nontoxic by-product (H₂O), as well as the reusability of the catalyst by simply washing with water and heating to 100°C.

Etherification Common ether synthesis typically involves treating an alkoxide anion with an alkyl halide, where a strong homogeneous base such as NaH deprotonates an alcohol giving an alkoxide, which attacks an electrophile such as methyl iodide. A NaOH base will suffice when using alcohols such as phenol and dialkyl sulfates are often used as electrophiles, leaving the sulfate ion in the waste stream. An interesting application recently published looks at transforming the potentially large quantity of glycerol by-product from the biodiesel industry (predicted to reach millions of tons) to useful glycerols ethers [51]. Amberlyst-35, Beta zeolite, Montmorillonite K10, Niobic acid, and p-toluene sulphonic acid (PTSA) were directly compared as acid catalysts, Fig. 11.

The stronger acid catalysts, beta zeolite and amberlyst-35, gave higher yields selectively to the mono benzyl-glycerol ether (around 55% and 38%,



Green Catalytic Transformations. Figure 11
Synthesis of Glycerol ethers

respectively, after only 2 h) due to their narrow porous channels and shape selectivity. Interestingly, montmorillonite K10, gave the overall highest conversion of the di benzyl-glycerol ether, ~56%, and also gave a high conversion of the mono benzyl-glycerol, of around 35%. Overall, the β -zeolite and montmorillonite K10 clay gave similar conversions of opposite selectivity, demonstrating that the observed acidity on the clay surface is only a part of its high activity. With a low stoichiometric ratio of benzyl alcohol to glycerol, 3:1, this reaction gives high conversions in short reaction times, allowing easy separation and reuse due to its heterogeneous nature and creates a valuable material from a potentially large future waste stream.

Alkylation The Friedel-Crafts alkylation is problematic when considering the frequency of its use in industry. The alkylation of aromatic rings typically involves reaction of pre-activated alkyl chain, such as 2-chloropropane, with powdered AlCl_3 catalyst or liquefied HF, which helps the alkyl halide to polarize, allowing π electrons from the aromatic benzene to attack the carbocation. Many clays have been investigated over past decades [52–54], and early work was focused on increasing the activity thorough ion exchange,

(for the most part the base clay does not show appreciable activity). Laszlo and Mathy [55] reported increased activity with exchanged clays when applied to various reactions such as the alkylation of aromatics with alkyl halides, alcohols, and alkenes. The test reaction was that of benzyl chloride and benzene, activity of the metal exchanged clays followed the following reactivity series [56], $\text{Fe}^{3+} > \text{Zn}^{2+} > \text{Cu}^{2+} > \text{Zr}^{4+}$, $\text{Ti}^{4+} > \text{Ta}^{5+} > \text{Al}^{3+} > \text{Co}^{2+} > \text{K10} > \text{Nb}^{5+}$. This not only showed an excellent alternative for alkylation catalysis, but more importantly demonstrated how support surfaces, whether clay, zeolite, or amorphous oxide, provide a very different environments for reacting species when compared to that of homogeneous solutions. For instance, the activity displayed by Zn is much higher than that of Al, where the opposite is true for the homogeneous chloride salt equivalents [19]. The range in activity is thought to be related to water coordinating to the metal center in a highly polarizing environment, exhibiting varying degrees of Bronsted acidity. A number of examples of solid acid catalysts and subsequent reactants/reagents are shown in Table 8 [57]. By far one of the greatest breakthroughs was that of the clayzic catalyst for benzylations, this will be discussed later in the supported reagents section.

Sulfonylations Use of $\text{SO}_3/\text{R}_2\text{SO}$ with H_2SO_4 is commonly in the synthesis of sulfones. These are often used in pharmaceuticals, polymers, and agrochemicals. Aside from the typical drawbacks associated with homogeneous acid catalysis, there can be problems with selectivity between para/ortho isomer in the sulfonylated of aromatic species. Both bronsted acidic and ion exchanged zeolites [58, 59] and clays have been researched as heterogeneous replacements, of which Fe^{+3} exchanged montmorillonite clay has shown high activity toward this type of transformation [60]. The activity of exchanged metals on montmorillonite follows $\text{Fe}^{3+} > \text{Zn}^{2+} > \text{Cu}^{2+} > \text{Al}^{3+} > \text{K10}$ in the sulfonylation of m-xylene and toluene-p-sulfonic anhydride [61]. A screen of various reagents using beta zeolite and Fe^{3+} montmorillonite can be seen in Table 9. The selectivity in the methanesulfonylation of toluene parallels that of AlCl_3 , achieving 33% selectivity toward the para position, 99% para selectivity is observed in the arenesulfonylation (TsCl). The activity

Green Catalytic Transformations. Table 8 Application of solid acids to alkylations

Reaction type	Catalyst	Reactant	Reagent
<i>Alkylation of aromatic hydrocarbons with</i>	Wyoming montmorillonite expanded with AlCl_3 and/or silanized with $(\text{EtO})_4\text{Si}$	Benzene	Ethane
<i>i) Olefins</i>	H-ZSM-5	Toluene	Ethane
<i>ii) Saturated hydrocarbons</i>	FeCl_3 doped montmorillonite K10	Benzene	Admantane
<i>iii) Alcohols</i>	ZnCl_2 clay, Al_2O_3 expanded smectite clay	Benzene, toluene naphthalene	Aliphatic alcohols cyclopentanol
	H-ZSM-5	Benzene	Ethanol
	Pillared montmorillonite, piller saponite, H-ZSM-5, HY	Toluene	Methanol
	H-ZSM-5	Naphthalene, methyl naphthalene	Methanol
<i>iv) Halide</i>	Transition metal salt deposited on montmorillonite K10	Benzene	Benzyl chloride
<i>Alkylation of phenols</i>	Activated clay	Phenol	Isobutene
	Bentonite		Oleic acid
	Supported/mixed oxides		Methanol
<i>Alkylation of aromatic amines</i>	Bentonite and Triton B	Diphenylamine	Styrene

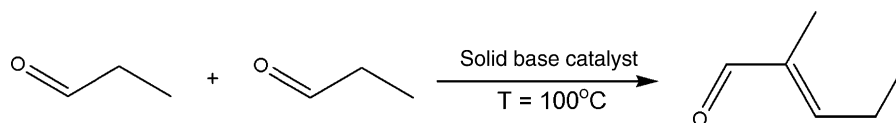
Green Catalytic Transformations. Table 9 Sulfonylation of various species using zeolite and montmorillonite

Catalyst	Sulfonylating agent	Yield (%)	Selectivity		
			<i>O</i>	<i>M</i>	<i>P</i>
Fe^{3+} montmorillonite	Ms_2O	81	47	19	34
Fe^{3+} montmorillonite	PhSO_2Cl	84	5	0	95
Fe^{3+} montmorillonite	TsCl	86	1	0	99
Fe^{3+} montmorillonite	TsOH	42	13	7	80
Fe^{3+} montmorillonite	Ts_2OH	84	11	6	83
Beta Zeolite	Ms_2O	78	42	21	37
Beta Zeolite	PhSO_2Cl	72	11	3	86
Beta Zeolite	TsCl	53	1	0	99
Beta Zeolite	TsOH^a	50	12	5	83
Beta Zeolite	Ts_2OH	88	14	6	80

^aStandard reaction time 6 h, TsOH reaction time 24 h in refluxing toluene

of the catalysts is thought to be to the mixed of Lewis and Bronsted acid catalysis. Reusability of the catalysts is found to be greater in systems that do not use Cl as an activating agent.

Aldol Condensation The use of clays for acid catalyzed reactions is well researched; however, little investigation (in comparison) is focused on their applications as a base. This largely stems from the



Green Catalytic Transformations. Figure 12

Synthesis of 2-methylpentanal

Green Catalytic Transformations. Table 10 Conversion of 2-methylpentanal

Catalyst	Conversion (%)	Selectivity (%)		
		2-methylpentanal	3-Hydroxy-2-methylpentanal	3-Pentanone
HT (1.5)	44	68	30	2
HT (2.0)	47	83	17	-
HT (2.5)	80	92	8	-
HT (3.0)	86	96	4	-
HT (3.5)	97	99	-	1

components looking to be replaced, in that, typically homogeneous acids, AlCl_3 , HCl , H_2SO_4 , tend to be more of a problem than, say, NaOH or KOH . Nevertheless similar problems still remain with stoichiometric quantities of waste twinned with disposal cost, as well as difficulty of product/reagent separation. A good example of utilizing basic properties of zeolites is given in the solvent free aldol condensation of propanol to 2-methylpentanal [62], Fig. 12. Propanol itself tends to be limited to solvent applications; alternatively, 2-methylpentenal has commercial importance in pharmaceuticals, fragrances, flavors, and cosmetics. With optimum reaction conditions using a homogeneous base such as NaOH , a high yield of 99% is obtained yet only a selectivity of 86% is achieved. As such, a number of solid bases have been investigated with hydrotalcite showing enhanced selectivity.

Unlike other materials, hydrotalcite (HT) has a typical composition of $\text{Mg}_6\text{Al}_2(\text{OH})_{16}(\text{CO}_3)_4(\text{H}_2\text{O})$, its name based on its resemblance to talc and high water content. The activity of the HT is found to vary with $\text{Mg}:\text{Al}$ ratio, from 1.5–3.5, directly proportional to the basicity of the material, Table 10. It is thought that two types of basic site exist in hydrotalcite, firstly a weak OH^- Bronsted and secondly a stronger O^- Lewis base [63].

Limitations There are few limitations of modern clay catalysis, most acid-treated clays show enhanced acidity in comparison to untreated and thermal decomposition of the clay structure can be avoided by pillaring the sheets. Increased selectivity is often seen and the ability of clay to ion exchange a variety of metals and make clay catalysis an efficient and green alternative to homogeneous activating agents.

Supported Reagents

Introduction

A supported reagent can be defined as a species that is supported, through either chemisorption or physisorption, onto an organic (such as ion exchange resin/carbon) or inorganic material. Typically supporting materials range from clays, to amorphous and structured silicas, alumina, and zeolites. Often the supported species is inorganic, such as precious metals, although there are a number of organic examples. These offer many advantages over their homogeneous equivalents:

- Reduced or nonexistent toxicity, leading to easier and safer handling
- Reusability

- Enhanced activity due to surface characteristic and active site distribution
- Increased selectivity though structural characteristic of supporting material
- Enantioselective potential

It should be noted, however, that complications related to poor diffusion, blocked pores due to substrates, potential poisoning of active species, and physical use on a large scale (fine silt and powders) sometimes create a few drawbacks.

Types of Support

Different supports have very different properties; however, for the majority of applications high surface areas of $>100 \text{ m}^2 \text{ g}^{-1}$ are considered advantageous, as this typically parallels concentration of active species. Depending on the reaction, in particular, the properties of supports can be tailored to suit, such as the acidity/basicity of the surface (the majority of applications utilized an acidic surface), pore size, and channel regularity. This can range from around 4 \AA for zeolitic supports to $>150 \text{ \AA}$ for structured silicas, temperature stability from $\sim 200^\circ\text{C}$ for standard acid-treated clays to $\sim 800^\circ\text{C}$ for amorphous silica (high temperature catalysts are often preferred for obvious reasons). Shape selectivity is often acquired with small/restrictive pores found in microporous silicas, zeolites, and pillared clays; however, diffusional problems and blockages can cause loss of activity over time. To overcome this problem many different species are supported on mesoporous supports, such as large pore SBAs, where pore size can reach around 200 \AA (some macroporous carbons can be above 500 \AA), although shape selectivity is nearly always lost at this size. Table 11 shows commonly used supports and corresponding surface areas [64]. Choice of support is an important consideration when designing clean synthetic routes. For instance, the acid properties of one support may destroy reagents whereas another may stabilize important intermediate species on the surface.

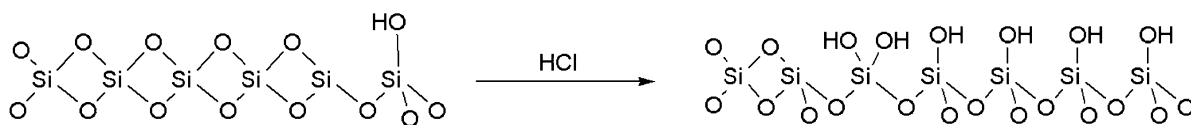
Preparation of Supported Species

Prior to loading active species on a support, certain pretreatments are commonly performed. Dehydration of the surface at around 120°C upward under vacuum, avoids hydrolysis complications and competitive

Green Catalytic Transformations. Table 11 Typical surface areas of commonly used supports

Support	Surface area m^2/g	Notes
Amorphous Silica Gel	300–600	Weakly acid surface due to isolated silanols, although is available in neutral and basic forms. Pores sizes typically range from 4 to 10 nm
Amorphous Alumina	100–300	Weakly basic, but through treatment and variation of surface groups can be acidic or neutral. Pore sizes similar to silica
Acid treated montmorillonites	50–300	K10 commonly used, give enhanced acid character compared with untreated montmorillonite clay. Able to swell and ion exchange with suitable solvent
Pillared montmorillonites	200–500	Stable at higher temperatures and can give enhanced shape selectivity
Structured Silica (HMS, SBA)	300–800	A wide choice of pore shapes and sizes available, allowing the supported to be tailored to fit certain syntheses. High temperature tolerance, easy to functionalise surface though SiOH silanols
Zeolites	300–600	Highly selective crystalline microporous materials. Able to exchanges ions to tune acidity type and strength

adsorptions onto binding sites. In the synthesis of $\text{AlCl}_3\text{-Al}_2\text{O}_3$, for example, AlCl_3 is readily hydrolysed unless the alumina is dried at least 550°C before impregnation. Another common technique used increases the number of available surface hydroxyls able to react on silica supports (often using HCl), effectively increasing the concentration of active species able to bind to the surface, Fig. 13. A number of preparation methods exist, depending on the end use of the catalysts, different methods are more suited than others.



Green Catalytic Transformations. Figure 13

Hydrolysis of siloxane bridges to give increased silanol quantity

Wet impregnation/Evaporation is a commonly used technique for loading species onto a support, due to its suitability for a wide range of reagents and supports. The method is based on dissolving the species of choice in a solvent of reasonable volatility, i.e., DCM, acetone, ethanol and stirring the solution with the support (typically 1–2 h). Once even wetting and diffusion into the pores has occurred the solvent can be evaporated *slowly*, ensuring even distribution of the species on the surface, deep in the porous network of the material. Post-treatment often involves calcining the resulting material, (anything from 200°C to 600°C), allowing chemisorption of the species onto the surface, referred to as thermal activation. *Precipitation* is generally used when there are issues with solubility, often with of metal salts. Deposition of the active species on the surface is achieved through either reaction with a second species to form an insoluble salt, cooling a hot solution containing the solute, or addition of cosolvent in which the active species is insoluble. *Adsorption* methodology simply involves stirring the support and reagent together (sometimes under reflux), yet requires a strong interaction between the surface and supporting species, aiming to achieve chemisorption if possible. After adsorption, suitable post treatment ensures the species is bound to the surface. Other methods have also been used such as *mixing/grinding* support and reagent together, in situ supporting (where the two solids are introduced directly into the reaction and the supported reagent is formed and reacts as the reaction proceeds) and ultrasound, these tend to be used less often.

Aside from post-modification, it is possible to synthesize materials with a species of choice available and evenly dispersed on the surface, known as sol-gel synthesis. The technique allows organic functionality to be incorporated into the structure and gives high surface area materials with high ratios of active species per gram. This methodology uses soluble (generally monomeric) precursors to the “support” and the supported

component, polymerising them together, often round templates to give highly structured materials. In addition, the organic species can often be reacted with many other substituents to tailor the functionality, amino functionalized silicas are particularly good at this.

Loading the Support

The dispersion of the active species becomes an important variable when physisorbing/chemisorbing a reactive species, such as a sulphonic acid or metal halide onto a surface. It should be noted that in other cases such as ion exchange of various cations with clays and zeolites, there are a defined number of sites that cannot be exceeded. Correct ratios of support to reagent can be estimated from the surface area and porosity of the material, optimal loadings tend to be between 0.5 and 2 mg per gram of support, aiming to achieve monolayer coverage. Porosimetry using N₂ adsorption onto the surface gives useful information when predicting potential concentration/loadings; however, problems can arise when using microporous supports, often is the case that N₂ can diffuse into pores that reagents cannot. As such, the exact requirement for various supports, active species and reactions, needs detailed investigation to achieve high efficiency. Overloading and underloading a support equally create a number of problems. Overloading can cause agglomeration of reagents leading to blocked pores, reducing the overall surface area available for catalysis and hindering diffusion. After monolayer dispersion is achieved, a second (bi-layer) can form, diluting the unique surface-reagent properties, where second and third layer species tend toward homogeneous characteristics. Contrastingly, underloading the support can cause lower catalytic activity, therefore requires increase quantities of the catalyst. With a lower concentration of active species, an increase in exposed surface is observed, which can be reactive enough to catalyze by-product formation.

Catalysis

There has been extensive research into supporting Bronsted and Lewis acids on different supports, resulting in many different catalyst properties and applications. Below is a sample out of a number of examples, showing the effect of different supports, supported species, and reactions, giving a much generalized view of the area. As with reactions previously discussed, through supporting active homogeneous species, the production of waste, risk factor, and cost is substantially reduced, giving high efficiency reactions with minimal impact on the environment and resources.

Bronsted Acid Catalysis

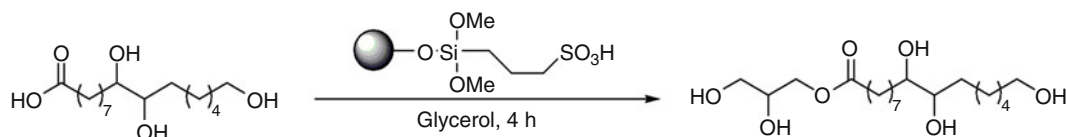
The acidity of supports, i.e., isolated silanols and alumina hydroxyls, tend to parallel organic acids, achieving a pK_a of around 4–6. This is of little use for the majority of acid catalyzed reactions and as such often requires a species of stronger acidity to be bound to the surface. Typically there are several types of Bronsted acidity observed in relation to supports, the original acidity from mixed oxides (silica and alumina hydroxyls), zeolitic acidity, activated/dissociated H₂O molecules usually coordinated with a Lewis acid metal on the surface and homogeneous Bronsted acids tethered or chemisorbed onto the support. The latter allows very high acid strength species to become reusable and safe to handle, below are a few examples.

Sulphonic Acid on Silica The hazards of using fuming acids on a large scale (as well as the subsequent tax deterrents on chemical waste) are well known and as a result there has been much investigation into supporting them on a variety of different media. One of the more “straight forward” supported acids is that of sulphuric acid/sulphonic acids supported on silica, demonstrating high activity and selectivity in a number

of transformations [65–73]. Jérôme et al. have shown that sulphonic acid supported on hexagonal mesoporous silica (HMS) gives selective esterification of multifunctional carboxylic acids with glycerol (a by-product of the biodiesel process) [74]. The reaction choice relates to its poor selectivity and rapid polymerisation to many undesired products, for instance the use of homogeneous p-toluene sulfonic acid in the esterification of glycerol with 16-hydroxyhexadecanoic acid gives a low yield of the amphiphilic monomer, only 35%. By anchoring the sulphonic species to HMS and SBA-15, an impressive increase in selectivity to 98% was observed, due to the restrictive nature of the pores. On application to other, more complex acids, high activity and selectivity remain consistent, Fig. 14.

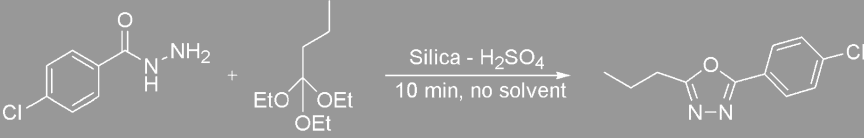
Other examples of sulphonic acids on silica gel include the synthesis of mono and disubstituted oxadiazoles from acyl hydrazides and orthoesters at room temperature [75]. Heterocycles are an important class of biologically active compounds often used as pharmaceutical intermediates. Out of a number of different solid acids investigated, sulphuric acid on silica gave high yields in as little as 10 min, Table 12. Interestingly, enhance activity was seen when compared to homogeneous H₂SO₄, showing how supported materials are often much more than the sum of their parts. Reuse of the catalyst showed that activity remained high even after five uses.

Shaabani [76] has also recently reported high yields using silica-H₂SO₄ in the synthesis of trisubstituted imidazoles by the one pot condensation of benzyl, benzoin, or benzylmonoxime with a substituted benzaldehyde and ammonium acetate. The synthesis used water as a solvent, whereas previously solvents such as methanol, ethanol, acetic acid, and DMSO were typically used. Excessive water in acid catalyzed reactions is known to deactivate acid centers; nevertheless yields of ~70% are achieved after 4–6 h.



Green Catalytic Transformations. Figure 14
Esterification using sulphonic acid on HMS

Green Catalytic Transformations. Table 12 Synthesis of 2-(4-Chloro-phenyl)-5-propyl-[1,3,4]oxadiazole [1]

	
Catalyst	Yield (%)
p-TsOH	42
NaHSO ₄	33
NaHSO ₃	28
H ₂ SO ₄	25
Montmorillonite K10	48
Silica sulphuric acid	94

Sulphated Zirconia Zirconia as a support has received attention in recent years due to its high activity toward hydrocarbon conversions. In particular, sulphated zirconia (SZ, $\text{SO}_4^{2-} - \text{ZrO}_2$) has shown catalytic potential toward light alkane isomerisation, as well as acylation, alkylation, cracking, and ring opening reactions [77–81]. As well as the many applications of mesoporous SZ to liquid phase reactions, standard microporous SZ is well suited to a number of vapor phase synthesis. It had previously been thought that the high activity of sulphated zirconia had been due to its super acidic properties, with an acid strength 1,000 times greater than homogeneous sulphuric acid [82]; however, more recent literature has shown that this is not the case [83]. FTIR spectroscopic titration with pyridine shows that surfaces activated between 500 and 600°C comprise of Bronsted acidity only [84], thought to be due to rapid adsorption of water, converting any Lewis acidity to Bronsted. Ratnam et al. recently reported of using SZ to catalyze the acylation of alcohols/phenols and amines, typically achieving yields of ~90% in 10 min [85].

Silica Gel as a Catalyst The use of silica as a support has been widely published, yet there has been little report of using silica gel alone as an acid catalyst. Amide synthesis has been a consistent problem area in green chemistry for many years. The majority of the problems stem from the used of stoichiometric quantities of activating agents used to create

a δ positive center on carboxylic acids, allowing nucleophilic attack from an amine. It should be noted that there are other type of reagents used for the synthesis of amides, such as nitriles, but the direct condensation of acid and amine is advantageous, due to their relatively low toxicity and by-product formation of water. The use of activating agent creates an equivalent of waste, for instance if an acid is activated to an anhydride, the second “low cost” acid is lost in the waste stream, excluding the proportion that has reacted with the amine, resulting in the formation of side products. There have been attempts at developing homogeneous catalysts, such as Yamamoto’s functionalised 3,4,5-trifluorophenyl boronic acid catalysts, though these are often limited to simple reactions, cannot be recovered and their synthesis is not straightforward. However, Clark et al. recently demonstrated the activity of calcined silica gel in the direct synthesis from carboxylic acids and amines [86]. Initially a number of supported metals, FeCl_3 and ZnCl_2 on silica and montmorillonite clay were investigated, but it was found that the silica gel alone gave the best activity. Unactivated, silica gel is a mild desiccant and does not typically display any catalytic activity towards organic reactions. On heating, a number of surface changes occur, around 120–200°C physisorbed water is lost, above this temperature the silanols start to dehydroxylated to give siloxane bridges up to around 1,000°C (when all of the surface silanols are lost to give quartz). It was found that activation at 700°C gave a hydrophobic

Green Catalytic Transformations. Table 13 Synthesis of amides using thermally activated silica gel

$\text{R}^1-\text{C}(=\text{O})\text{OH} + \text{H}_2\text{N}-\text{R}^2 \xrightarrow[\text{Toluene}^{[b]}]{\text{K60}^{[a]}} \text{R}^1-\text{C}(=\text{O})\text{NHR}^2$			
R ¹ [c]	R ²	Uncatalysed yield	Isolated yield (%) ^a
C ₆ H ₅	C ₆ H ₅	0	47 ^b
C ₆ H ₅ CH ₂	C ₆ H ₅	10	81
CH(CH ₃)C ₆ H ₅	C ₆ H ₅	4	72 ^b
CHClCH ₃	C ₆ H ₅	4	70
OCH ₂ C ₆ H ₄ Cl	C ₆ H ₅	0	73
CH ₂ C ₆ H ₅	C ₆ H ₄ Cl	0	48
CH ₂ C ₆ H ₅	C ₆ H ₄ (CH ₃) ₂	0	38 ^b
CH ₂ C ₆ H ₅	C ₄ H ₈	2	89
CH ₂ CH ₃	CH ₂ CH ₂ CH ₂ CH ₃	89 ^c	98 ^d

^aAfter 24 h reflux in toluene, 20%wt silica^bUsing 50% wt silica^cAfter 12 h^dAfter 24 h

surface with mild Bronsted acidity and high activity toward amidation. A number of different reactions are shown below, Table 13.

Despite the use of high loadings of silica with particularly difficult syntheses, such as benzoic acid and aniline, the initial low cost, low toxicity and simple preparation and use makes the synthesis a substantial improvement on previous catalysts. Multiple reuse with reactivation (burning out organic buildup) is possible, without loss of activity. With a high atom economy of 93% and an E-factor of less than 1, the synthesis allows high conversions without the use of toxic and highly corrosive activating agents.

Lewis Acid Catalysis

Lewis acidity is often based on integrated surface cations acting as acid centers, where the strength largely depends on the surrounding environment. However, as previously discussed with zeolites and clays, water can coordinate to these acid centers and give Bronsted acidity, effectively “diluting” the Lewis acid effect. By tethering homogeneous species such as ZnCl₂ and FeCl₃ to various supports, increased Lewis acidity is

observed. An excellent example of this is in the development and use of EnvirocatsTM.

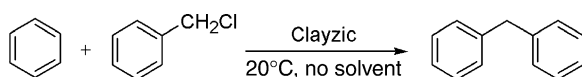
EnvirocatsTM Designed as a series of commercially available clay and alumina-based catalysts, EnvirocatsTM were developed by collaboration between the University of York and Contract Chemicals Ltd. (CCL). They comprise of a metal salt supported on a high surface area inorganic support and have a range of applications, Table 14. Although they are often term “simple” in that the species being supported is not a complicated structure, the mechanism by which they catalyze many different reactions is complex [87].

A good example of their enhanced activity is of the benzylation of benzene and benzylchloride with EPZ10, also known as clayzic, Fig. 15. The diphenylmethane product is often used as a precursor to many pharmaceuticals, making this particular reaction of commercial interest. As with many Lewis acid catalyzed reactions, AlCl₃ was formerly the homogeneous catalyst of choice, exhibiting little selectivity and giving high ratios of polybenzylated product, unless very large excesses of benzene are used. Clayzic, on

Green Catalytic Transformations. Table 14 EnvirocatsTM and common uses

Catalyst	Supported species	Type of reactions	Acidity
EPZG	Iron/Clay ^a	Friedel-Crafts Benzoylations, some acylations such as etherification	Contains both Bronsted and Lewis acid sites
EPZ10	ZnCl ₂ /K10 Clay	Friedel-Crafts alkylation of aromatics, Benzylations	Very strong Lewis acid, few weakly Bronsted acidic sites
EPZE	ZnCl ₂ /Clay	Friedel-Crafts Sulfonylations and some Benzoylations	Prodominatly strong Lewis acid with some strong Bronsted acids. Different preparation method to EPZ10
EPIC	Phosphoric acid/Clay	Esterifications, general Bronsted acid catalyzed reactions	Exclusively strong Bronsted acidity
EPAD	Cr(VI)/alumina	Oxidations	Cr(VI) oxidant center, not suitable for used with peroxide – causes leaching of the metal

^aAcid treated montmorillonite clay

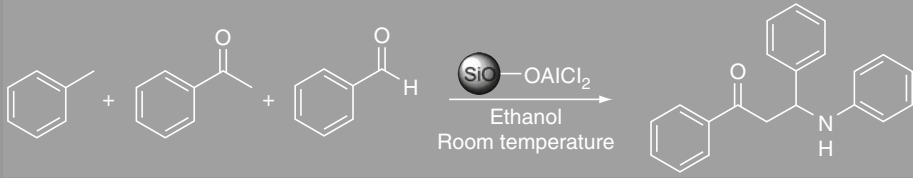
**Green Catalytic Transformations. Figure 15**
Benzylation of benzene and benzylchloride

the other hand, can be used in similar conditions to AlCl₃, yet gives extraordinarily high yields of ~75% after 4 h at room temperature, even with a minimal excess of benzene. In addition, the catalyst can be simply filtered off, washed and reused many times over, showing an excellent example of green chemistry being integrated into industry.

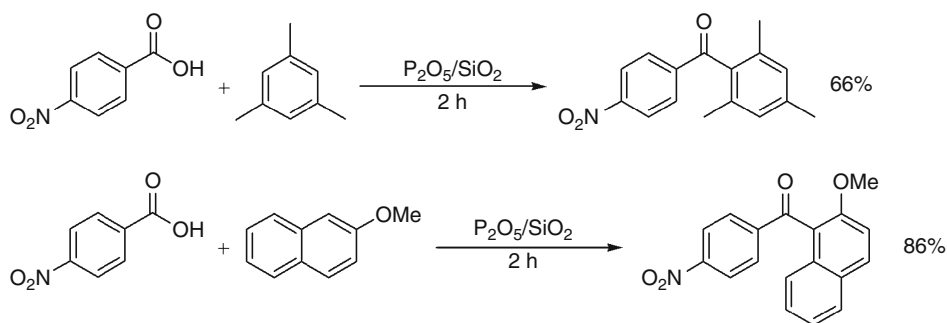
AlCl₃ on Silica Aluminum chloride is a widely used strong, in expensive homogeneous Lewis acid. There are, however, many problems associated with its use, including the necessitated destruction of the acid with water to aid separation of products, resulting high volumes of corrosive and toxic waste and formation of by-products due to its inherent activity. There have been a number of reports detailing the immobilization of aluminum chloride on a variety of supports to overcome these problems. Typical procedures involve dissolving the AlCl₃ in an organic solvent and slowly reacting this with the silica surface, theoretically creating a SiOAlCl₂ species capable of both Bronsted and Lewis acidity [86]. The resulting material has been applied to a range of reactions with success, early syntheses focus on the Friedel-Crafts acylation. Li et al. discusses the use of AlCl₃-silica in a one pot, three

component Mannich type reaction, Table 15 [88]. Activating agents such as HCl, InCl₃, Y(OTf)₃, Zn(BF₄)₂, Bi(OTf)₃, PS-SO₃H, phosphorodiamidic acid, dodecylbenzene sulfonic acid have been reported, suffering from the usual homogeneous drawbacks. Interestingly, the reaction of aniline, benzaldehyde, and acetophenone to give a β-aminocarbonyl (used in numerous pharmaceutical syntheses and natural products) showed that homogeneous AlCl₃ gave lower yields when compared with the supported species. The optimum reaction conditions also add to the green credentials of this synthesis, the best solvent out of a range including, MeCN, CH₂Cl₂, THF, and toluene has been found to be ethanol, with a reaction temperature of ~25°C.

Supported Lewis Acids in Acylation Freidel-Crafts reactions are a key target for green chemistry due to their frequent use in chemical industry and large volumes of waste produced. Two main problems areas of the synthesis are the initial activation to an acyl chloride, with subsequent generation of HCl and the use of a strong Lewis acid to cleave the Cl from the acyl chloride to produce an acylium ion, enabling nucleophilic attack from π electrons on the aromatic ring. Early developments focused on the direct replacement of the Lewis acid with a heterogeneous equivalent, examples include Ga₂O₃ and In₂O₃ supported on Si-MCM-41 [89, 90] and InCl₂, GaCl₃ and ZnCl₂ supported on both montmorillonite clay and Si-MCM-41 [91, 92]. Despite high yields, the catalyst

Green Catalytic Transformations. Table 15 Mannich type synthesis using aniline, acetophenone, and benzaldehyde


Catalyst	Time (hours)	Yield (%)
FeCl ₃ – SiO ₂	20	0
ZnCl ₂ – SiO ₂	20	0
AlCl ₃	20	0
SiO ₂	20	0
AlCl ₃ /SiO ₂ dry mix	5	74
AlCl ₂ – SiO ₂	5	93

**Green Catalytic Transformations. Figure 16**

Acylation using P₂O₅/SiO₂. (*Where solvent is 1,2-dichloroethane under reflux, % given is conversion)

required pre-activation to the acyl chloride, creating stoichiometric quantities of HCl waste. Further improvements have developed allowing anhydrides to be used as starting materials, still creating equivalents of waste, yet not as corrosive as previous mineral acid waste [36, 93–96]. However, recent publications discuss the ideal scenario, use of carboxylic acid starting material producing only water as a by-product [97, 98], P₂O₅/SiO₂ in particular has shown potential. Application to a range of aromatic acylations has given high yields (Fig. 16), on average ~20% higher when compared to the homogeneous P₂O₅ alone [99].

The inexpensive nature of the catalyst, simple preparation, and ease of handling, as well as the use of

unactivated carboxylic acids, producing only water by-product, make this synthesis improved over previous homogeneous species, as well as a number of heterogeneous catalysts.

Solid Base Catalysis

As mentioned before, the push for research into “clean” and reusable heterogeneous base catalysts has not been as pronounced as that for solid acids; however there is still a wide range of basic catalysts available. Many supporting materials can be thermally treated to exhibit basic properties (alumina), though they are weak and often species with increased basicity are supported. The basicity is usually the result of basic

Green Catalytic Transformations. Table 16 Supported basic species [64]

Solid base	Notes
Alumina – Na	Superbasic
Alumina – KOH	Widely used for many years
Alumina – KF	Widely used with variable basicity
Alumina – CsF	Little advantage over cheaper alumina-KF
Alumina – AlPO_4	Used in solventless reactions
Silica – Na	Superbasic
Silica – K	Superbasic
Silica – NaOH	Simplest form of basic silica
Zeolite- M^+	Can be very basic depending on alkali metal
Xonotlite-KOtBu	Unusual supports – similar to alumina-KF
MgO-K/Na	Lower surface area than alumina and silica

hydroxyl groups, oxygen ions bound to the surface or basic ions bound to the surface. Table 16 gives an overview of commonly used solid bases.

Basic Species Supported on Alumina (KF/Alumina)

First discussed by Clark [100], potassium fluoride supported on alumina (KF/Alumina) has been acknowledged as an extremely useful catalyst able to promote a variety of base catalyzed reactions [101]. The strength and nature of the basic sites on the surface has been a subject of much discussion over the years, but it has been widely accepted that the basic strength can vary from weak-moderate to superbasic depending on the preparation (drying condition of alumina post impregnation) and subsequent exposure to CO_2 , for example. The active species is thought to relate to the “hard” fluoride anion as opposed to oxygen anions. It has been observed that when K_2CO_3 and KOH are supported on alumina, the resulting activity is much lower than that of KF and K_3AlF_6 species (formed at high loadings of KF, itself exhibiting no activity) on the alumina surface, suggesting that the presence of fluorine is essential to its activity [102]. Interestingly, unlike other fluoride ions sources, there has never been any

report of KF-alumina exhibiting any nucleophilicity, even in simple substitutions. Another peculiarity of the material is that high activity is increasingly observed when KF loadings are higher than a monolayer equivalent, at low loadings no crystallinity can be observed through X-ray diffraction. Further identification of the F-species has been researched using [21] F NMR, yet the results have proved somewhat controversial [103]. KF-alumina remains one of the most important base catalysts due to its wide range of applications to many syntheses such as, isomerisations [103–105], condensations [102], Knoevenagel [106], cyclization [107], as well as C–N [108], C–O [109–111], S–C [112–114] and Si–O [115, 116] bond forming reactions. In particular, its application to Michael additions has received much attention [117–127], a typical example of this is in the addition of nitroalkanes to electron deficient alkenes [128]. The use of homogeneous catalysts (such as DBU) is often problematic due to the nitro acting as a leaving group, yielding nitric acid and unsaturated alkene. In addition, withdrawing groups on the α , β positions of the alkene make this a poor nucleophile. However, the use of KF-alumina as a catalyst gives impressive results compared to a range of alternative basic catalysts, Table 17.

Temperature control allows high selectivity between potential products, where room temperature conversions give the Michael addition adduct. A high E:Z stereoselective ratio of 95:5, is consistently found for the formation of the unsaturated 1,4 dicarbonyl derivatives from primary nitroalkanes. Overall the high activity, reusability, selectivity, and solvent-less nature of reactions utilizing KF/Alumina make it an excellent alternative to classic homogeneous bases.

Basic Amine Species Supported on Silica

Aside from basic metals, there are a range of supported organic species able to provide basicity, of which amine functionality has attracted much interest [129]. The amines in particular vary widely, from imines, phenolates [130] to dialkylaminopyridine type species [131]. Hagiwara et al. discusses the use of amino functionalised silica gel in the Knoevenagel reaction [132]. Out of a range of different silicas of varying morphology, (i.e., pellets/powders) as well as different amines, 3-Aminopropyl supported on powdered silica

Green Catalytic Transformations. Table 17 Michael addition of methyl 4-nitrobutanoate with dimethyl maleate [102]

Catalyst	Yield of route A ^c (%)
KF/Basic Alumina	80
KF/Basic Alumina	79 ^d
KF/Neutral Alumina	71
Basic Alumina	4
Hydrotalcite MG 70	7
Hydrotalcite MG 50	6
Amberlyst A-21	25
Amberlyst A-27	21
Silica supported-1,5,7-triazabicyclo-[4.4.0]dec-5-ene (TBD)	36
N,N-Diethylpropylamine supported on amorphous silica (KG-60-NEt ₂)	19

^aCatalysed with KF/Alumina at 55°C^bCatalysed using KF/Alumina at room temperature^cAfter 7 h, solventless reaction conditions^dYield of route B

has been found to be the most active. Typical yields of a range of different functionalized benzaldehydes reacted with ethyl cyanoacetate are in the region 90–99% in between 2–5 h. In addition, water is able to be used as a solvent, thought to allow both substrates to achieve close proximity to each other on the reverse phase silica gel, increasing reaction efficiency. This demonstrates how organic functionalized solid bases have many advantages over their inorganic equivalents. Often they are mild in reactivity, reducing potential side product formation, have high reproducibility, with flexibility in design allowing the catalyst to be “tailored to fit” a wide range of applications.

Future Directions

The value of catalysis in chemistry has been well known for many years with application in larger scale, petrochemistry especially important and well established in industry. This was emphasized by the emergence in the second half of the twentieth century of more selective,

typically zeolitic catalysts in continuous vapor phase processes that helped establish the petrochemical and commodity chemical industries as the engines of the great manufacturing industries of the world. An increasing awareness of the need to reduce the environmental impact of chemical manufacturing has caused a much wider range of industrial sectors to seek process improvements in chemical, pharmaceutical and polymer manufacturing. More efficient and selective processes that give less waste and greater resource efficiency and the avoidance of large quantities of often hazardous traditional reagents including acids, bases, and stoichiometric oxidants are important goals. Here, the transfer of long lifetime, selective, and reusable catalytic technologies from large-scale petrochemical processes to smaller scale, fine, and speciality (including pharmaceutical) chemical manufacturing has become and continues to be very important. It is needed to quickly move away from old, admittedly reliable but often dangerous and almost always very wasteful chemistry in all sectors of manufacturing.

In research, it is required to work ever harder to find effective replacements for the great reagents of the twentieth century – aluminum chloride, chromic acid, hydrogen fluoride, caustic soda, and many others – all cheap, versatile in numerous reactions, and readily available yet also all hazardous, unselective, and leading to more waste than product especially after work-up; progress to date has been limited and most processes continue little changed from decades in the past. The use of heterogeneous catalysts is required to extend the use of continuous production and intensive processing so as to get away from inefficient and inherently risky batch manufacturing. Catalysis should be the norm in all chemical manufacturing and not the exception, and the catalysts themselves must have verifiably sound lifecycles and low environmental footprints including ease of recovery and reuse especially when precious metals are employed.

Bibliography

- Clark JH (1995) Chemistry of waste minimization. Chapman and Hall, Cambridge
- Clark JH, Rhodes CN (2000) Clean synthesis using porous inorganic solid catalysts and supported reagents. Royal Society of Chemistry, Cambridge
- Rabo JA, Schoonover MW (2001) Early discoveries in zeolite chemistry and catalysis at Union Carbide and follow-up in industrial catalysis. *Appl Catal A* 222:261
- Nagy JB, Bodart P, Hannus I, Kirics I (1998) Synthesis, characterisation and use of zeolitic materials. DecaGen Ltd., Hungary, p165
- Sing KSW, Everett DH, Haul RAW, Moscou L, Pierotti RA, Rouquérol J, Siemieniowska T (1985) Reporting physisorption data for gas/solid systems with special reference to the determination of surface area and porosity. *Pure Appl Chem* 57:603
- Kissin YV (2001) Chemical mechanisms of catalytic cracking over solid acidic catalysts: alkanes and alkenes. *Catal Rev* 43(1):85
- Zhuang JQ et al (2004) Solid-state MAS NMR studies on the hydrothermal stability of the zeolite catalysts for residual oil selective catalytic cracking. *J Catal* 228(1):234
- Caeiro G, Magnoux P, Lopes JM, Ribeiro FR, Menezes SMC, Costa AF, Cerqueira HS (2006) Stabilization effect of phosphorus on steamed H-MFI zeolites. *Appl Catal A* 314(2):160
- Bao X et al (2005) Enhancement on the hydrothermal stability of ZSM-5 zeolites by the cooperation effect of exchanged lanthanum and phosphoric species. *J Mol Struct* 737(2–3):271
- Blasco T, Corma A, Martínez-Triguero J (2006) Hydrothermal stabilization of ZSM-5 catalytic cracking additives by phosphorus addition. *J Catal* 237(2):267
- Ding W et al (2007) Understanding the enhancement of catalytic performance for olefin cracking: hydrothermally stable acids in P/HZSM-5. *J Catal* 248(1):20
- Barros ZS, Zotin FMZ, Henriques CA (2007) Conversion of natural gas to higher valued products: light olefins production from methanol over ZSM-5 zeolites. *Stud Surf Sci Catal* 167:255
- Lu R, Cao Z, Liu X (2008) Catalytic activity of phosphorus and steam modified HZSM-5 and the theoretical selection of phosphorus grafting model. *J Nat Gas Chem* 17(2):142
- Gao X et al (2009) Modification of ZSM-5 zeolite for maximizing propylene in fluid catalytic cracking reaction. *Catal Commun* 10(14):1787
- Brouwer DM, Hogeveen H (1972) Electrophilic substitutions at alkanes and in alkylcarbonium ions. *Prog Phys Org Chem* 9:179
- Guisnet M, Andy P, Boucheffa Y, Gnep NS, Travers C, Benazzi E (1998) Selective isomerization of n-butenes into isobutene over aged H-ferrierite catalyst: nature of the active species. *Catal Lett* 50:159
- Santacesaria E, Di Serio M, Cozzolino M, Tesser R (2004) DGMK-conference C4/C5-hydrocarbons: routes to higher value-added products, Munich
- Seo G et al (1996) Skeletal isomerization of 1-butene over ferrierite and ZSM-5 zeolites: influence of zeolite acidity. *Catal Lett* 36(3–4):249
- Mériaudeau P, Tuan VA, Le NH, Szabo G (1997) Selective isomerization of n-Butene into isobutene over deactivated H-Ferrierite catalyst: further investigations. *J Catal* 169(1):397
- Asensi MA, Martínez A (1999) Selective isomerization of n-butenes to isobutene on high Si/Al ratio ferrierite in the absence of coke deposits: implications on the reaction mechanism. *Appl Catal A* 183:155
- Wichterlova B et al (1999) Effect of bronsted and lewis sites in ferrierites on skeletal isomerization of n-butenes. *Appl Catal A* 182(2):297
- Auerbach SM, Carrado KA, Dutta PK (2003) Handbook of zeolite science and technology. Marcel Dekker, p 481
- Shouro D et al (2000) Mesoporous silica FSM-16 catalysts modified with various oxides for the vapor-phase Beckmann rearrangement of cyclohexanone oxime. *Appl Catal A* 198(1):275–282
- Chaudhari K et al (2002) Beckmann rearrangement of cyclohexanone oxime over mesoporous Si-MCM-41 and Al-MCM-41 molecular sieves. *J Mol Catal A Chem* 177(2):247
- Zhang Y et al (2005) Beckmann rearrangement of cyclohexanone oxime over H β zeolite and H β zeolite-supported boride. *Catal Commun* 6:53
- Dai LX et al (1997) Development of advanced zeolite catalysts for the vapor phase Beckmann rearrangement of cyclohexanone oxime. *Appl Surf Sci* 121/122: 335
- Misono M, Inui T (1999) New catalytic technologies in Japan. *Catal Today* 51:369
- Izumi Y et al (2007) Development and Industrialization of the Vapor-Phase Beckmann Rearrangement Process. *Bull Chem Soc Jpn* 80(7):1280–1287

29. Roffia P, Leofanti G, Cesana A, Mantegazza M, Padovan M, Petrini G, Tonti S, Gervasutti P (1990) Cyclohexanone ammoximation: a break through in the 6-caprolactam production process. *Stud Surf Sci Catal* 55:43
30. Palkovits R, Schmidt W, Ilhan Y, Erdem-S-enatlar A, Schüth F (2009) Crosslinked TS-1 as stable catalyst for the Beckmann rearrangement of cyclohexanone oxime. *Microporous Mesoporous Mater* 117:228
31. Kumar R, Rao GN, Ratnasamy P (1989) Influence of the pore geometry of medium pore zeolites ZSM-5, -22, -23, -48 and -50 on shape selectivity in reactions of aromatic hydrocarbons. *Stud Surf Sci Catal* 49:1141
32. Sotelo JL, Uguina MA, Valverde JL, Serrano DP (1993) Kinetics of toluene alkylation with methanol over magnesium-modified ZSM-5. *Ind Eng Chem Res* 32:2548
33. Climent MJ, Corma A, Vely A (2004) Synthesis of hyacinth, vanilla, and blossom orange fragrances: the benefit of using zeolites and delaminated zeolites as catalysts. *Appl Catal A* 263(2):155
34. Climent MJ, Corma A, Garcia H, Guil-Lopez R, Iborra S, Fornés V (2001) Acid–base bifunctional catalysts for the preparation of fine chemicals: synthesis of jasminaldehyde. *J Catal* 197(2):385
35. Climent MJ, Corma A, Vely A, Susarte M (2000) Zeolites for the production of fine chemicals: synthesis of the fructose fragrance. *J Catal* 196(2):345
36. Kantam ML et al (2005) Friedel–Crafts acylation of aromatics and heteroaromatics by beta zeolite. *J Mol Catal A Chem* 225(1):15
37. Andy et al (2000) Acylation of 2-methoxynaphthalene and isobutylbenzene over zeolite beta. *J Catal* 192(1):215
38. Heinichen HK, Holderich WF (1999) Acylation of 2-methoxynaphthalene in the presence of modified zeolite HBEA. *J Catal* 185(2):408
39. Casagrande M, Storaro L, Lenarda M, Ganzerla R (2000) Highly selective Friedel–Crafts acylation of 2-methoxynaphthalene catalyzed by H-BEA zeolite. *Appl Catal A* 201(2):263
40. Bejblova M, Zilkova N, Cejka J (2008) Transformations of aromatic hydrocarbons over zeolites. *Res Chem Intermed* 34(5–7):439
41. Guggenheim S, Martin RT (1995) Definition of clay and clay mineral: joint report of the IUPAC nomenclature and CMS nomenclature committees. *Clays Clay Miner* 43(2):255
42. Velde B, Meunier A (2008) The origin of clay minerals in soils and weathered rocks. Springer, Berlin Heidelberg, Chapter 1
43. Clark JH, Rhodes CN (2000) Clean synthesis using porous inorganic solid catalysts and supported reagents. *R Soc Chem, Chapter 3 clay materials*, p 38
44. Duc M et al (2005) Sensitivity of the acid–base properties of clays to the methods of preparation and measurement: 1. Literature review. *J Colloid Interface Sci* 289:139
45. Brown DR, Rhodes CN (1997) Bronsted and Lewis acid catalysis with ion-exchange clays. *Catal Lett* 45(1–2):35
46. Ravichandran J, Lakshmanan CM, Sivasankar B (1996) Acid activated montmorillonite and vermiculite clays as dehydration and cracking catalysts. *React Kinet Catal Lett* 59(2):301
47. Ravichandran J, Sivasankar B (1997) Properties and catalytic activity of acid-modified montmorillonite and vermiculite. *Clays Clay Miner* 45(6):854
48. Klopogge JT, Duong LV, Frost RL (2005) A review of the synthesis and characterization of pillared clays and related porous materials for cracking of vegetable oil to produce biofuels. *Environ Geol* 47:967
49. Reddy CR et al (2007) Surface acidity study of Mn²⁺–montmorillonite clay catalysts by FT-IR spectroscopy: correlation with esterification activity. *Catal Commun* 8:241
50. Reddy CR et al (2004) Synthesis of phenylacetates using aluminium-exchanged montmorillonite clay catalyst. *J Mol Catal A Chem* 223(1):117
51. da Silva et al (2009) Etherification of glycerol with benzyl alcohol catalyzed by solid acids. *J Braz Chem Soc* 20(2):201
52. Salmon M, Zavala N, Martinez M, Miranda R, Cruz R, Cardenas J, Gavino R, Cabrera A (1994) Cyclic and linear oligomerization reaction of 3, 4, 5-trimethoxybenzyl alcohol with a bentonite-clay. *Tetrahedron Lett* 35(32):5797
53. Sabu KR, Sukumar R, Lalithambika M (1993) Acidic properties and catalytic activity of natural kaolinitic clays for Friedel–Crafts alkylation. *Bull Chem Soc Jpn* 66:3535
54. Okada S, Tanaka K, Nakadaira Y, Nakagawa N (1992) Selective Friedel–Crafts alkylation on a vermiculite, a highly active natural clay mineral with Lewis acid sites. *Bull Chem Soc Jpn* 65:2833
55. Laszlo P, Mathy A (1987) Catalysis of Friedel–Crafts alkylation by a montmorillonite doped with transition-metal cations. *Helv Chim Acta* 70(3):577
56. Clark JH, Macquarrie DJ (1997) Heterogeneous catalysis in liquid phase transformations of importance in the industrial preparation of fine chemicals. *Org Process Res Dev* 1:149
57. Narayanan S, Deshpande K (2000) Aniline alkylation over solid acid catalysts. *Appl Catal A* 199(1):1
58. Smith K, Ewart GM, El-Hiti GA, Randlesb KR (2004) Study of regioselective methanesulfonylation of simple aromatics with methanesulfonic anhydride in the presence of zeolite catalysts. *Org Biomol Chem* 2:3150
59. Laidlaw P, Bethell D, Brown SM, Watson G, Willock DJ, Hutchings GJ (2002) Sulfonylation of substituted benzenes using Zn-exchanged zeolites. *J Mol Catal A Chem* 178(1):205
60. Choudary BM, Chowdari NS, Kantam ML, Kannan R (1999) Fe (III) exchanged montmorillonite: a mild and ecofriendly catalyst for sulfonylation of aromatics. *Tetrahedron Lett* 40:2859
61. Choudary BM, Chowdari NS, Kantam ML (2000) Friedel–Crafts sulfonylation of aromatics catalysed by solid acids: an ecofriendly route for sulfone synthesis. *J Chem Soc Perkin Trans* 1:2689
62. Sharma SK, Parikh PA, Jasra RV (2007) Solvent free aldol condensation of propanal to 2-methylpentenal using solid base catalysts. *J Mol Catal A Chem* 278(1):135

63. Cavani F, Trifiro F, Vaccari A (1991) Hydrotalcite-type anionic clays: preparation, properties and applications. *Catal Today* 11:173
64. Clark JH (1994) *Catalysis of organic reactions by supported inorganic reagents*. VCH Publishers Inc, New York
65. Salehi P, Ali Zolfigol M, Shirini F, Baghbanzadeh M (2006) Silica sulfuric acid and silica chloride as efficient reagents for organic reactions. *Curr Org Chem* 10(17):2171
66. Li Z, Liu J, Gong X, Mao X, Sun X, Zhao Z (2008) Silica sulfuric acid-catalyzed expeditious environment-friendly hydrolysis of carboxylic acid esters under microwave irradiation. *Chem Pap* 62(6):630
67. Mobinikhaledi A, Foroughifar N, Fard MAB, Moghanian H, Ebrahimi S, Kalhor M (2009) Efficient one-pot synthesis of polyhydroquinoline derivatives using silica sulfuric acid as a heterogeneous and reusable catalyst under conventional heating and energy-saving microwave irradiation. *Synth Commun* 39:1166
68. Zarei A, Hajipour AR, Khazdooz L, Mirjalili BF, Chermahini AN (2009) Rapid and efficient diazotization and diazo coupling reactions on silica sulfuric acid under solvent-free conditions. *Dyes Pigm* 81(1):240
69. Chen X, She J, Shang Z, Wu J, Zhang P (2009) Room-temperature synthesis of pyrazoles, diazepines, β -enaminones, and β -enamino esters using silica-supported sulfuric acid as a reusable catalyst under solvent-free conditions. *Synth Commun* 39:947
70. Wang Y, Yuan Y, Guo S (2009) Silica sulfuric acid promotes aza-Michael addition reactions under solvent-free condition as a heterogeneous and reusable catalyst. *Molecules* 14:4779
71. Shobha D, Chari MA, Mukkanti K, Ahn KH (2009) Silica gel-supported sulfuric acid catalyzed synthesis of 1, 5-benzodiazepine derivatives. *J Heterocycl Chem* 46(5):1028
72. Li J, Meng X, Bai B, Sun M (2010) An efficient deprotection of oximes to carbonyls catalyzed by silica sulfuric acid in water under ultrasound irradiation. *Ultrason Sonochem* 17:14
73. Yang J, Dang N, Chang Y (2009) Silica sulfuric acid as a recyclable catalyst for a one-pot synthesis of α -aminophosphonates in solvent-free conditions. *Lett Org Chem* 6(6):470
74. Karam A, Gu Y, Jérôme F, Douliez J, Barrault J (2007) Significant enhancement on selectivity in silica supported sulfonic acids catalyzed reactions. *Chem Comm* 22:2222
75. Dabiri M et al (2007) Silica sulfuric acid: an efficient and versatile acidic catalyst for the rapid and ecofriendly synthesis of 1,3,4-oxadiazoles at ambient temperature. *Synth Commun* 37:1201
76. Shaabani A, Rahmati A (2006) Silica sulfuric acid as an efficient and recoverable catalyst for the synthesis of trisubstituted imidazoles. *J Mol Catal A Chem* 249(1):246
77. Reddy BM, Patil MK (2009) Organic syntheses and transformations catalyzed by sulfated zirconia. *Chem Rev* 109(6):2185
78. Reddy BM, Sreekanth PM, Lakshmanan P (2005) Sulfated zirconia as an efficient catalyst for organic synthesis and transformation reactions. *J Mol Catal A Chem* 237(1):93
79. Deutsch J, Trunschke A, Müller D, Quaschnig V, Kemnitz E, Lieske H (2004) Acetylation and benzylation of various aromatics on sulfated zirconia. *J Mol Catal A Chem* 207(1):51
80. Deutsch J, Prescott HA, Müller D, Kemnitz E, Lieske H (2005) Acylation of naphthalenes and anthracene on sulfated zirconia. *J Catal* 231(2):269
81. Zane F, Melada S, Signoretto M, Pinna F (2006) Active and recyclable sulphated zirconia catalysts for the acylation of aromatic compounds. *Appl Catal A* 299:137
82. Hino M, Arata K (1980) Synthesis of solid superacid catalyst with acid strength of $H_0^* = -16.04$. *J Chem Soc, Chem Commun* (18):851
83. Paukshtis EA, Shmachkova VP, Kotsarenko NS (2000) Acidic properties of sulfated zirconia. *React Kinet Catal Lett* 71(2):385
84. Clark JH (2002) Solid acids for green chemistry. *Acc Chem Res* 35:791
85. Ratnama KJ, Reddy RS, Sekhar NS, Kantama ML, Figueras F (2007) Sulphated zirconia catalyzed acylation of phenols, alcohols and amines under solvent free conditions. *J Mol Catal A Chem* 276(1):230
86. Comerford JW, Clark JH, Macquarrie DJ, Breeden SW (2009) Clean, reusable and low cost heterogeneous catalyst for amide synthesis. *Chem Commun* 14(18):2562
87. Clark JH, Macquarrie DJ (2002) *Handbook of green chemistry and technology*. Blackwell Science Ltd, Chapter 13, Green Catalysts for Industry
88. Li Z et al (2007) Silica-supported aluminum chloride: a recyclable and reusable catalyst for one-pot three-component Mannich-type reactions. *J Mol Catal A Chem* 272(1):132
89. Choudhary VR, Jana SK, Kiran BR (2000) Highly active Si-MCM-41-supported Ga_2O_3 and In_2O_3 catalysts for friedel-crafts-type benzylation and acylation reactions in the presence or absence of moisture. *J Catal* 192(2):257
90. Choudhary VR, Jana SK (2002) Acylation of aromatic compounds using moisture insensitive mesoporous Si-MCM-41 supported Ga_2O_3 catalyst. *Synth Commun* 32(18):2843
91. Choudhary VR, Jana SK, Patil NS (2001) Acylation of benzene over clay and mesoporous Si-MCM-41 supported $InCl_3$, $GaCl_3$ and $ZnCl_2$ catalysts. *Catal Lett* 76(3):235
92. Choudhary VR, Patil KY, Jana SK (2004) Acylation of aromatic alcohols and phenols over $InCl_3$ /montmorillonite K-10 catalysts. *J Chem Sci* 116(3):175
93. Derouane EG, Dillon CJ, Bethell D, Derouane-Abd Hamid SB (1999) Zeolite catalysts as solid solvents in fine chemicals synthesis: 1. catalyst deactivation in the Friedel–Crafts acylation of anisole. *J Catal* 187(1):209
94. Derouane EG, Crehan G, Dillon CJ, Bethell D, He H, Derouane-Abd Hamid SB (2000) Zeolite catalysts as solid solvents in fine chemicals synthesis: 2. competitive adsorption of the reactants and products in the Friedel–Crafts acylations of anisole and toluene. *J Catal* 194(2):410
95. Yadav GD, George G (2006) Friedel–Crafts acylation of anisole with propionic anhydride over mesoporous superacid catalyst UDCaT-5. *Microporous Mesoporous Mater* 96(1–3):36

96. Ishitani H, Naito H, Iwamoto M (2008) Friedel-Crafts acylation of anisole with carboxylic anhydrides of large molecular sizes on mesoporous silica catalyst. *Catal Lett* 120(1–2):14
97. Sarvari MH, Sharghi H (2005) Solvent-free catalytic Friedel-Crafts acylation of aromatic compounds with carboxylic acids by using a novel heterogeneous catalyst system: p-toluenesulfonic acid/graphite. *Helv Chim Acta* 88:2282
98. Waghlikar SG, Niphadkar PS, Mayadevi S, Sivasanker S (2007) Acylation of anisole with long-chain carboxylic acids over wide pore zeolites. *Appl Catal A* 317(2):250
99. Zarei A, Hajipour AR, Khazdooz L (2008) Friedel-Crafts acylation of aromatic compounds with carboxylic acids in the presence of P_2O_5/SiO_2 under heterogeneous conditions. *Tetrahedron Lett* 49:6715
100. Clark JH (1980) Fluoride ion as a base in organic synthesis. *Chem Rev* 80:429
101. Blass BE (2002) KF/Al_2O_3 mediated organic synthesis. *Tetrahedron* 58:9301
102. Handa H, Baba T, Sugisawa H, Ono Y (1998) Highly efficient self-condensation of benzaldehyde to benzyl benzoate over KF -loaded alumina. *J Mol Catal A Chem* 134(1–3):171
103. Kabashima H, Tsuji H, Nakatab S, Tanaka Y, Hattori H (2000) Activity for base-catalyzed reactions and characterization of alumina-supported KF catalysts. *Appl Catal A* 194–195:227
104. Tsuji H, Kabashima H, Kita H, Hattori H (1995) Thermal activation of KF /alumina catalyst for double bond isomerization and Michael addition. *React Kinet Catal Lett* 56(2):363
105. Kochkar H, Clacens JM, Figueras F (2002) Isomerization of styrene epoxide on basic solids. *Catal Lett* 78(1–4):91
106. Nakano Y, Niki S, Kinouchi S, Miyamae H, Igarashi M (1992) Knoevenagel reaction of malononitrile with acetone followed by double cyclization catalyzed by KF -coated alumina in aqueous solution. *Bull Chem Soc Jpn* 65(11):2934
107. Wang WC, Wang D, Forray C, Vaysse PJJ, Branchekeo TA, Gluchowski C (1994) A convenient synthesis of 2-amino-2-oxazolines and their pharmacological evaluation at cloned human α adrenergic receptors. *Bioorg Med Chem Lett* 4(19):2317
108. Yamawaki J, Ando T, Hanafusa T (1981) N-Alkylation of amides and N-heterocycles with potassium fluoride on alumina. *Chem Lett* 1143–1146
109. Yamawaki J, Ando T (1980) Potassium fluoride on alumina as a base for crown ether synthesis. *Chem Lett* 9(5):533–536
110. Sawyer JS, Schmittling EA (1993) Synthesis of diaryl ethers, diaryl thioethers, and diarylamines mediated by potassium fluoride-alumina and 18-crown-6. *J Org Chem* 58(12):3229
111. Yadav VK, Kapoor KK (1996) KF adsorbed on alumina effectively promotes the epoxidation of electron deficient alkenes by anhydrous t-BuOOH. *Tetrahedron* 52:3659
112. Moghaddam FM, Bardajee GR, Veranlou ROC (2005) KF/Al_2O_3 -mediated Michael addition of thiols to electron-deficient olefins. *Synth Commun* 35(18):2427
113. Villemin D, Alloum AB (1992) Potassium fluoride on alumina: an easy synthesis of 4-alkylidene-2-thione-1, 3-oxathiolanes from α -acetylenic alcohols. *Synth Commun* 22:1351
114. Villemin D, Hachemi M, Lalaoui M (1996) Potassium fluoride on alumina: synthesis of O-aryl N, N-dimethylthiocarbamates and their rearrangement into S-aryl N, N-dimethylthiocarbamates under microwave irradiation. *Synth Commun* 26(13):2461
115. Kawanami Y, Yuasa H, Toriyama F, Yoshida S, Baba T (2003) Addition of silanes to benzaldehyde catalyzed by KF loaded on alumina. *Catal Commun* 4:455
116. Baba T, Kato A, Yuasa H, Toriyama F, Handa H, Ono Y (1998) New Si-C bond forming reactions over solid-base catalysts. *Catal Today* 44:271
117. Clark JH, Cork DG, Robertson MS (1983) Fluoride ion catalysed Michael reactions. *Chem Lett* 12(8):1145
118. Campelo JM, Climent MS, Marinas JM (1992) Michael addition of nitromethane to 3-buten-2-one catalyzed by potassium fluoride supported on Al_2O_3 , ZnO , SnO_2 , sepiolite, $AlPO_4$, $AlPO_4-Al_2O_3$ and $AlPO_4-ZnO$. *React Kinet Catal Lett* 47:7
119. Kabashima H, Tsuji H, Shibuya T, Hattori H (2000) Michael addition of nitromethane to α , β -unsaturated carbonyl compounds over solid base catalysts. *J Mol Catal A Chem* 155(1–2):23
120. Wang SH, Wang XS, Shi DQ, Tu SJ (2003) Michael addition reaction of dimedone and chalcone catalyzed by KF/Al_2O_3 . *Chin J Org Chem* 23(10):1146
121. Figueras F et al (2004) Effect of the support on the basic and catalytic properties of KF . *J Catal* 221(2):483
122. Tian DB, Zhu J, Zhu JF, Shi YX, Wang JT (2004) Michael addition of alkyl amine to α , β -unsaturated carbonyl compounds catalyzed by KF/Al_2O_3 . *Chin Chem Lett* 15(8):883
123. Moghaddam FM, Bardajee GR, Taimoory SMD (2006) KF/Al_2O_3 mediated aza-Michael addition of indoles to electron-deficient olefins. *Lett Org Chem* 3(2):157
124. Wang X, Quan Z, Wang JK, Zhang Z, Wang M (2006) A practical and green approach toward synthesis of N3-substituted dihydropyrimidinones: using Aza-Michael addition reaction catalyzed by KF/Al_2O_3 . *Bioorg Med Chem Lett* 16(17):4592
125. Lenardão EJ, Ferreira PC, Jacob RG, Perin G, Leiteb FPL (2007) Solvent-free conjugated addition of thiols to citral using KF /alumina: preparation of 3-thioorganilycitronealls, potential antimicrobial agents. *Tetrahedron Lett* 48:6763
126. Lenardão EJ, Trecha DO, Ferreira PC, Jacob RG, Perin G (2009) Green Michael addition of thiols to electron deficient alkenes using KF /alumina and recyclable solvent or solvent-free conditions. *J Braz Chem Soc* 20(1):93
127. Clark JH, Farmer TJ, Macquarrie DJ (2009) The derivatization of bioplatfrom molecules by using KF /Alumina catalysis. *ChemSusChem* 2(11):1025
128. Ballinia R, Palmieri A (2006) Potassium fluoride/basic alumina as far superior heterogeneous catalyst for the chemoselective conjugate addition of nitroalkanes to electron-poor alkenes having two electron withdrawing groups in α - and β -positions. *Adv Synth Catal* 348(10):1154
129. Macquarrie DJ (2009) Organically modified micelle templated silicas in green chemistry. *Top Catal* 52(12):1640

130. Utting KA, Macquarri DJ (2000) Silica-supported imines as mild, efficient base catalysts. *New J Chem* 24:591
131. Motokura K, Tomita M, Tada M, Iwasawa Y (2009) Michael reactions catalyzed by basic alkylamines and dialkylamino-pyridine immobilized on acidic silica-alumina surfaces. *Top Catal* 52:579
132. Isobe K, Hoshi T, Suzuki T, Hagiwara H (2005) Knoevenagel reaction in water catalyzed by amine supported on silica gel. *Mol Divers* 9(4):317

Green Chemistry and Chemical Engineering, Introduction

The goal of Green Chemistry and Chemical Engineering is to minimize waste, totally eliminate the toxicity of waste, minimize energy use, and utilize green energy (solar thermal, solar electric, wind, geothermal, etc.) – that is, non fossil fuel. Clearly, fossil fuels have their own waste and toxicity problems even though usually remote from the site of chemical production.

The objective in preparing this section is to provide a significant sampling of the scientific and engineering basis of green chemistry and engineering with specific processes as examples. There are nine detailed entries written at a level for use by university students through practicing professionals. For ease of use by students, each entry begins with a glossary of terms, while at an average length of 20 print pages each, sufficient detail is presented for utilization by professionals in government, universities, and industry. The reader is also directed to closely related sections in our *Encyclopedia: Solar Thermal Energy* (see the entry, ► [Solar Energy in Thermochemical Processing](#)); *Hydrogen Production Science and Technology* (see the entry, ► [Photo-catalytic Hydrogen Production](#) and also ► [Hydrogen via Direct Solar Production](#)); *Solar Radiation* (see the entry on ► [Photosynthetically Active Radiation: Measurement and Modeling](#)); *Geothermal Power Stations*; and the section on *Batteries*.

Each of the entries is summarized below.

► [Gas Expanded Liquids for Sustainable Catalysis](#) – A gas-expanded liquid (GXL) phase is generated by dissolving a compressible gas such as CO₂ or a light olefin into the traditional liquid phase at mild pressures (tens of bar) (). When CO₂ is used as the expansion gas, the resulting liquid phase is termed a CO₂-expanded

liquid or CXL. GXLs combine the advantages of compressed gases such as CO₂ and of traditional solvents in an optimal manner. GXLs retain the beneficial attributes of the conventional solvent (polarity, catalyst/reactant solubility) but with higher miscibility of permanent gases (O₂, H₂, CO, etc.) compared to organic solvents at ambient conditions and enhanced transport rates compared to liquid solvents (ii, iii, iv, v). The enhanced gas solubility's in GXLs have been exploited to alleviate gas starvation (often encountered in homogeneous catalysis with conventional solvents), resulting in a 1–2 orders of magnitude greater rates than in neat organic solvent or scCO₂. *Environmental advantages* include substantial replacement of organic solvents with environmentally benign CO₂. *Process advantages* include reduced flammability due to CO₂ presence in the vapor phase and milder process pressures (tens of bar) compared to scCO₂ (hundreds of bar). GXLs thus satisfy many of the attributes of an ideal alternative solvent.

► [Green Catalytic Transformations](#) – A heterogeneous catalyst is a catalytically active species that is in a different phase to the reagents within a reaction system, more often than not a solid in either liquid or vapor phase synthesis. There are a number of green chemistry related advantages of using heterogeneous catalysts as opposed to the homogeneous equivalents: Safety – An important consideration in the practice of green chemistry. Heterogeneous catalysts often tend to be environmentally benign and easy/safe to handle. This is due to the active species being adhered to a support, (often forming a powder) essentially reducing its reactivity with the surrounding environment; Reusability – Due to the difference in phases, the catalyst is simply filtered off (through centrifugation on industrial scales) and reactivated for reuse many times over; Activity – In many cases increased activity is observed when supporting an active homogeneous species on a support, due to the complex but unique surface characteristics found with a variety of different supports; and Selectivity – heterogeneous catalysts can give an increased degree of selectivity in reaction pathway. This can simply be a consequence of adsorption of substrates and the consequential restricted freedom of movement of the reacting molecules. However, there are a handful of disadvantages. The quantity of solid catalyst required is often higher than that of the

homogeneous equivalent, due to the lower concentration on the support surface, (reusability makes this less of an issue though). Blocking of the pores/support channels can occur with narrow pores sizes and reduce efficiency over time in some liquid phase reactions; nevertheless, this is often twinned with high stereoselectivity.

► **Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design** – Application of green metrics forces the precise itemization of what constitutes waste so that targeted reductions of these components can be made. The mass of waste of any chemical reaction is the sum of the masses of unreacted starting materials, byproducts produced as a mechanistic consequence of making the desired target product, side products produced from competing side reactions other than the intended reaction, reaction solvent, all work-up materials, and all purification materials used. Simple first generation waste reduction strategies target the last three items in the list since they contributed the bulk of the overall mass of waste. Waste reduction strategies targeting the first three items are based on synthesis design and are necessarily more challenging to implement. The connecting green metrics are atom economy (AE), environmental E-factor (E), and reaction mass efficiency (RME). These metrics and applications are described in detail in this entry.

► **Green Chemistry with Microwave Energy** – An alternative heating technique using microwaves is useful for targeted energy introduction directly into polar reactants in chemical syntheses and transformations. This entry summarizes noteworthy greener methods that use microwaves that have resulted in the development of sustainable synthetic protocols for drugs and fine chemicals. Microwave assisted organic transformations are presented such as: solid-supported reagents based processes; greener reaction media including aqueous, ionic liquid, and solvent-free for the synthesis of various heterocycles; and oxidation-reduction and also coupling reactions.

► **Nanotoxicology in Green Nanoscience** – The unique properties that make nanomaterials an attractive technology (surface chemistry, surface area, size, shape, core material functionalization, aggregation, etc.) may also contribute to novel biological effects as a result of nanomaterial exposure. Toxicology will play an important role in elucidating the mechanisms of

those interactions. This entry contains an exploration of the role of toxicology in implementing green nanoscience, the methodology of incorporating nanotoxicology in order to directly and indirectly address the principles of green chemistry in green nanoscience, and the importance of utilizing robust models for nanotoxicity testing.

► **Organic Batteries** – Development of sustainable processes for energy storage and supply is one of the most important worldwide concerns today. Primary batteries, such as alkaline manganese and silver oxide batteries produce electric current by a one-way chemical reaction and are not rechargeable and hence useless for reversible electricity storage. Portable electronic equipment, electric vehicles, and robots require rechargeable secondary batteries. Li-ion, lead acid, and nickel-metal hydride batteries are generally used at the present to power them. Solar cells and wind-power generators expect a parallel use of rechargeable batteries for leveling and preserving their generated electricity. Ubiquitous electronic devices such as integrated circuit smart cards and active radio-frequency identification tags need rechargeable batteries that are bendable or flexible and environmentally benign for durability in daily use. Designing of soft portable electronic equipment, such as rollup displays and wearable devices, also require the development of flexible batteries. It is essential to find new, low-cost, and environmentally benign electroactive materials based on less-limited resource for electric energy storage and supply.¹ Reversible storage materials of electric energy or charge that are currently under use in electrodes of rechargeable batteries are entirely inorganic materials, such as Li ion-containing cobalt oxide, lead acid, and nickel-metal hydride.

► **Oxidation Catalysts for Green Chemistry** – Catalysis is at the heart of Green Chemistry as it is the means to increase efficiency and efficacy of chemical and energy resources while promoting environmental friendliness and intensifying time and cost savings in chemical synthesis. A catalyst's function simply is to provide a pathway for chemicals (reactants) to combine in a more effective manner than in its absence. In the absence of a catalyst, heat is usually the way to overcome the energy barrier but this increases energy consumption and often results in unwanted side

reactions. A catalyst cannot make an energetically unfavorable reaction occur or change the chemical equilibrium of a reaction because the catalyzed rate of both the forward and the reverse reactions are equally affected. Oxidation processes are used for odor control, bleaching of pulp for paper production, wastewater treatment, disinfection, bulk and specialty chemical production, aquatics and pools, food and beverage processing, cooling towers, agriculture/farming, and many others. There are a great many oxidation reactions that are catalytically driven. Such reactions are presented in this entry with emphasis on implementation of green strategies.

► **New Polymers, Renewables as Raw Materials** –

Recent advances in genetic engineering, composite science, and natural fiber development offer significant opportunities for developing new, improved materials from renewable resources that can biodegrade or be recycled, enhancing global sustainability. A wide range of high-performance, low-cost materials can be made using plant oils, natural fibers, and lignin. These materials have economic and environmental advantages that make them attractive alternatives to petroleum-based materials.

► **Supercritical Carbon Dioxide (CO₂) as Green Solvent** – A supercritical fluid (SCF) is created when the temperature and pressure are higher than its critical values. Therefore, CO₂ becomes supercritical when its temperature and pressure are higher than 31.1°C (critical temperature, T_c) and pressure 7.38 MPa (critical pressure, P_c). SCFs have many unique properties, such as strong solvation power for different solutes, large diffusion coefficient comparing with liquids, zero surface tension, and their physical properties can be turned continuously by varying the pressure and temperature because the isothermal compressibility of SCFs is very large, especially in the critical region. These unique properties of SCFs lead to great potential for the development of innovative technologies. Besides these common advantages of SCFs, supercritical CO₂ (scCO₂) has some other advantages, such as nontoxic, nonflammable, chemically stable, readily available, cheap, and easily recyclable, and it has easily accessible critical parameters. Therefore, scCO₂ can be used as green solvent in different fields. The basic properties of scCO₂ and its applications in extraction and fractionation, chemical reactions, polymeric

synthesis, material science, supercritical chromatography, painting, dyeing and cleaning, and emulsions related with CO₂, are discussed in this entry.

Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design

JOHN ANDRAOS

Department of Chemistry, York University, Toronto, ON, Canada

Article Outline

Glossary
 Definition of the Subject and Its Importance
 Introduction
 Overview of Algorithms
 Strategy Metrics
 An Example: Sertaline Plans
 Summary of Optimization
 Thresholds and Probabilities
 Future Directions
 Bibliography

Glossary

Atom economy Ratio of molecular weight of target molecule to sum of molecular weights of reactants assuming a balanced chemical equation.

B/M Ratio of sum of number of target bonds made in a synthesis plan to total number of reaction steps.

By-product of a reaction A product formed in a reaction between reagents as a direct mechanistic consequence of producing the target product assuming a balanced chemical equation that accounts for the production of that target product.

Degree of asymmetry A parameter determined from the shape of a synthesis tree diagram for a synthesis plan that describes the degree of skewness of a triangle whose vertices are the target product node, the origin, and the node for the last reagent along the ordinate axis.

Degree of convergence A parameter determined from the shape of a synthesis tree diagram for a

synthesis plan that describes the ratio of the angle subtended at the actual product node vertex to that at a product node vertex corresponding to the hypothetical case of all reaction substrates in a plan reacting in a single step.

E-factor with respect to molecular weight (E_{mw})

Ratio of sum of molecular weights of by-products in a reaction to molecular weight of target product in a given reaction or synthesis plan.

E-factor with respect to mass (E_m) The ratio of mass of total waste from all sources to mass of target product collected in a given reaction or synthesis plan.

E_{kernel} Contribution to the total or overall E-factor with respect to mass from reaction by-products, reaction side products, and unreacted starting materials.

E_{excess} Contribution to the total or overall E-factor with respect to mass from excess reagents.

$E_{auxiliaries}$ Contribution to the total or overall E-factor with respect to mass from auxiliary materials such as reaction solvents, work-up, and purification materials.

$E_{overall}$ (or E_m) Ratio of mass of total waste generated from all sources to mass of target product collected for a given reaction or synthesis plan.

f^* Fractional kernel waste contribution from target bond forming reactions.

$f(sac)$ Molecular weight fraction of reagents that absolutely do not end up in the final target molecule structure.

$f(nb)$ Fraction of number of reaction stages that do not form target bonds.

Hypsicity index A parameter that tracks the oxidation numbers of atoms in target bond forming reactions over the course of a synthesis plan relative to their values in the target molecule.

I Number of input materials in a synthesis plan.

Kernel mass of waste Mass of all reagents used in a synthesis reaction or plan minus the mass of target product collected.

Kernel reaction mass efficiency The ratio of the mass of target product collected to the sum of the stoichiometric masses of all reagents used in a chemical reaction or synthesis plan.

μ_1 , First molecular weight moment Parameter that describes the degree of building up going on from

the reagent molecules toward the target molecule over the course of a synthesis plan.

Mass intensity Ratio of mass of all materials used to mass of target product collected in a given chemical reaction or synthesis plan.

Material recovery parameter A parameter that describes the mass consumption of all solvents and auxiliary materials used in carrying out a given chemical reaction that may be potentially recoverable by recycling.

M Number of reaction steps in a synthesis plan.

N Number of reaction stages in a synthesis plan.

Radial hexagon A diagram that depicts the overall atom economy, overall reaction yield along the longest branch, overall kernel reaction mass efficiency, molecular weight fraction of reagents in whole or in part that end up in the target molecule, fraction of the kernel waste contribution from target bond forming reactions, and the degree of convergence for a given synthesis plan as a hexagon for easy visualization of the synthesis performance.

Radial pentagon A diagram that depicts the reaction yield, atom economy, stoichiometric factor, material recovery parameter, and overall reaction mass efficiency for a given chemical reaction as a pentagon for easy visualization of the reaction performance.

Raw material cost Sum of the costs of all reagents in a synthesis plan calculated using the basis scale in moles of the target product, assuming balanced chemical equations for all reactions in the plan and taking account of excess reagent consumption.

Reaction step Refers to interval between a given isolated intermediate and the next consecutive isolated intermediate in a synthesis plan.

Reaction yield Ratio of moles of target product to moles of limiting reagent for a given balanced chemical equation.

Sacrificial reactions Nonproductive reactions in a synthesis plan that do not form target bonds appearing in the target product structure.

Side product of a reaction A product formed in a reaction between reagents, usually undesired, that arises from a competing reaction pathway other than the one that produces the intended target product and its associated by-products.

Stoichiometric factor (SF) A parameter that describes the total amount of excess reagents used in a given chemical reaction relative to the amounts prescribed by its stoichiometrically balanced chemical equation.

Stoichiometric coefficient An integer appearing before the chemical formula for a reagent in a balanced chemical equation.

Synthesis tree A diagram that describes all features of a synthesis plan, including number of steps, number of stages, number of branches, number of intermediates, number of reagents used, and molecular weights of all chemical species.

Target bond map A chemical structure drawing of the target product of a synthesis plan showing the target bonds made, the step numbers of each target bond, and the set of atoms that correlate directly with the corresponding reagents that ended up in the target molecule.

Target bond forming reactions Productive reactions in a synthesis plan that result in the formation of bonds that appear in the structure of the target molecule.

Total (overall) mass of waste The sum of all masses of materials used in synthesis plan minus the mass of the target product collected.

Total (overall) reaction mass efficiency The ratio of the mass of target product collected to the sum of the masses of all reagents, solvents, and auxiliaries used in a given reaction or synthesis plan.

UDI A parameter that tracks the “oxidation length” traversed over the course of a synthesis plan equal to the sum of all the “ups” and “downs” in a hypsicity profile or bar graph beginning with the zeroth reaction stage.

Definition of the Subject and Its Importance

Over the last 2 decades, the topic of “green metrics” has grown rapidly in conjunction with the field of green chemistry. Green metrics promise to provide a rigorous, thorough, and quantitative understanding of material, energy, and cost efficiencies for individual chemical reactions and synthesis plans. Indeed, before the advent of green chemistry, good synthetic strategy and elegance were ill-defined, yet intuitive concepts couched less in quantitative terms and more by subjective ones. The quest for a reliable method of measuring

material efficiency or “greenness” of a chemical reaction, synthesis, or process is of fundamental importance in the field of organic synthesis when various routes to a given target molecule are considered for selection. Such a method should be robust in its application to any kind of reaction or plan regardless of complexity. It should standardize the ranking of efficiencies of synthesis plans in an unambiguous and unbiased way. It should be used as a powerful tool in weeding out bad-performing plans quickly and identifying promising candidate plans at the drawing board stage guided by thorough literature investigations even before embarking on any experiment in the laboratory. Above all, it should facilitate the decision-making process for both chemists and non-chemists by allowing on-the-spot precise identification of potential bottlenecks in plans from easy-to-read graphics and by pointing chemists in the right direction for further optimization with confidence. In the long term, with enough examples and pattern recognition among as many plans as possible, it should be possible to understand why “good” plans are good and offer insights into what good strategy entails and how to parameterize it. *The key take home message is that optimization is an iterative exercise that compares the performances of a set of plans to a common target according to some criteria and that true optimization is achieved when the best possible values of material efficiency and synthetic elegance metrics coincide in the same plan.*

Introduction

In the period 2005–2007, algorithms were introduced to assess the material efficiencies of any reaction and any synthesis plan, and they delivered on all of the points mentioned above [1–6]. Their first and most important triumph was that they provided simple connecting relationships between key green metrics that previously were considered as independent entities. The basis of this discovery was the long forgotten law of conservation of mass pronounced by Antoine Lavoisier in 1775 that allows one to write out complete balanced chemical equations for every reaction in a synthesis plan with appropriate stoichiometric coefficients for all reactants and *all* by-products in addition to the target product of interest. The skill of balancing chemical equations has been resurrected after a long period of dormancy in the field of synthetic organic

chemistry by the need to precisely identify waste composition and quantify waste production, which are essential if any real progress can be made in the newly emerging field of green chemistry whose chief aim is to minimize waste. In fact, before the advent of green metrics, waste was considered as one big lump of unwanted material, which was calculated simply as the difference between the mass of all input materials used and the mass of the desired output material. Application of green metrics forces the precise itemization of what constitutes waste so that targeted reductions of these components can be made. The mass of waste of any chemical reaction is the sum of the masses of unreacted starting materials, by-products produced as a mechanistic consequence of making the desired target product, side products produced from competing side reactions other than the intended reaction, reaction solvent, all work-up materials, and all purification materials used. Simple first-generation waste reduction strategies target the last three items in the list since they contributed the bulk of the overall mass of waste. Waste reduction strategies targeting the first three items are based on synthesis design and are necessarily more challenging to implement. The connecting green metrics are atom economy (AE) [7–9], environmental E-factor (E) [10–12], and reaction mass efficiency (RME) [13, 14]. All other metrics appearing in the literature [15–21] that pertain to material efficiency can be expressed in terms of these three and are therefore redundant. RME and E describe identically the same thing from the positive and negative points of view, respectively. Experimental atom economy for a reaction is just another synonym for the kernel RME (reaction yield times atom economy), mass intensity (MI) is just $E + 1$, and mass index is the reciprocal of RME. Prior arguments that AE had a lower usefulness than RME and E were dispelled by the simple algebraic connecting relationships found; namely, $AE = 1/(1 + E_{mw})$ and $RME = 1/(1 + E)$, where E_{mw} and E are the E-factors based on reagent molecular weight and mass, respectively. In fact, AE is of fundamental importance in the molecular design phase of synthesis planning even before carrying out any experiments as it describes the intrinsic waste produced in a reaction or plan due to by-products arising out of the nature of the chemical reaction in question. If significant by-products are known to result from reactions in a plan at the outset,

then it is inevitable that this will have an additive and cumulative effect on the production of overall waste.

This chapter summarizes the salient features of these algorithms, one for individual reactions and the other for synthesis plans. The reader is referred to prior references by the author for derivations. Numerical computations are greatly facilitated using spreadsheet programs which have been disclosed previously in a detailed investigation of the material efficiencies of the synthesis plans for oseltamivir phosphate a neuraminidase inhibitor used to treat influenza, particularly H5N1 and H1N1 [22]. Here, we illustrate how these algorithms may be implemented by showing complete worked out examples for the analysis of seven literature synthesis plans for the antidepressant pharmaceutical sertraline [23–29]. This compound is unique because its published plans show a steady progression in optimization over the period 1984–2006. This is a very rare situation in the literature where the timeline of publications shows a steady upward rise in synthesis efficiency.

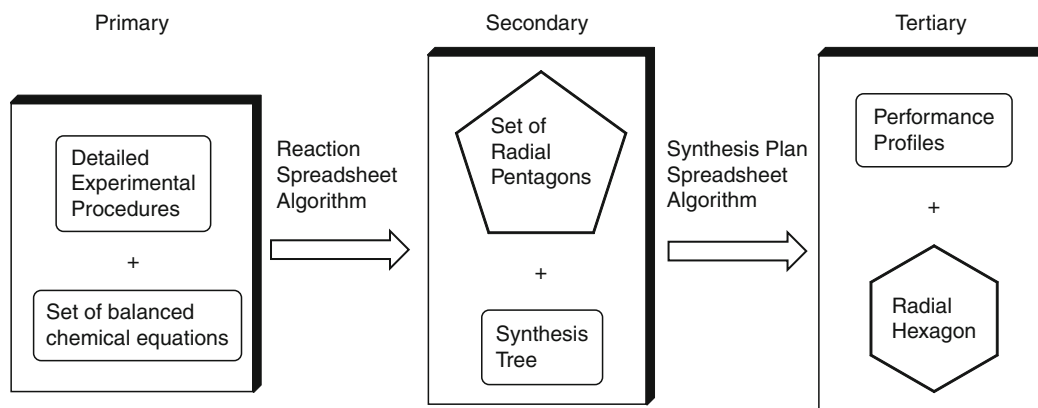
Overview of Algorithms

The main theme and sequence of algorithms used in carrying out green metrics calculations on a given synthesis plan are shown in Fig. 1. The first thing is to access the best available literature procedures that disclose the experimental write-ups for the recipes for each reaction in the synthesis plans to the target compound of interest. Next, fully balanced chemical equations for all reactions are written down where all by-products arising from the intended mechanisms are identified and stoichiometric coefficients for all reactants and products are assigned. From these data, radial pentagons are constructed according to the method described in [5] to obtain diagrams that show the AE, reaction yield, stoichiometric factor (SF), and material recovery parameter (i.e., reaction solvent, work-up, and purification material demand) for each reaction according to the master equation given by Eq. 1:

$$RME = (\varepsilon)(AE)\left(\frac{1}{SF}\right)(MRP) \quad (1)$$

where

ε is the reaction yield with respect to the limiting reagent in a reaction given by the mole ratio of the target product collected and of the limiting reagent.



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 1
Paradigm flowchart for determining green metrics for any synthesis plan

AE is the reaction atom economy given by the ratio of the molecular weight of the target product to the sum of molecular weights of all reagents in the balanced chemical equation.

SF is the stoichiometric factor, taking into account the use of excess reagents, and is given by

$$SF = 1 + \frac{\sum \text{excess masses of all reagents}}{\sum \text{stoichiometric masses of all reagents}} \quad (2)$$

MRP is the material recovery parameter, taking into account the use of auxiliary materials, and is given by

$$MRP = \frac{1}{1 + \frac{(\epsilon)(AE)(c + s + \omega)}{m_p(SF)}} \quad (3)$$

where m_p is the mass of the collected target product, c is the mass of catalyst used, s is the mass of reaction solvent, and ω is the mass of all other auxiliary materials used in the work-up and purification phases of the reaction.

For ease of computation, the four factors in Eq. 1 are written as fractions between 0 and 1 instead of as percentages. Eq. 1 basically shows how RME is factored or partitioned into its four constituent contributors. The corresponding radial pentagon gives a clear visual representation of Eq. 1 and tells a chemist which of the parameters may be the bottlenecks in attenuating the RME value for a given reaction. The MRP factor is the strongest attenuator because the total mass of auxiliary materials far outweighs the masses of all

reagents used. Hence, this was the area that was targeted in first-generation waste reduction when the 12 principles of green chemistry were announced. The radial pentagon algorithm incorporates two check calculations, one for RME using Eq. 1 and by simply dividing the mass of collected target product by the sum of the masses of all the input materials used in the procedure, and the other for SF using Eq. 2 and from $SF = (\epsilon)(AE)/(RME)$ where RME, in this case, refers to the reaction mass efficiency calculated when all auxiliary materials are assumed to be recovered, that is, not committed to waste so that MRP is set to 1.

A huge bonus from the radial pentagon analysis is that it picks up all errors that appear in literature procedures. The most common are incorrect reporting of reaction yields, masses of ingredients, and moles of ingredients. Unfortunately, there is less care in the reporting of these key parameters by authors in experimental sections of research papers than expected. Moreover, reviewers and editors of journals do not routinely check for the accuracy of these details, particularly if they are buried in supplementary material, as they naively assume that authors have assumed that responsibility. Patent documents are notorious for such errors as they are often made with intent. Practicing chemists know all too well that such mistakes come to light when someone else embarks on duplicating the reported experiment. The biggest drawback in determining the true RME for a reaction is the lack of full disclosure of all masses of auxiliary materials used in a given

procedure. Of particular note are the amounts of solutions used in work-up washes and extractions, mass of drying agent used, and masses of chromatographic materials (solid supports and eluents). If these problems remain unchecked, the successful wide implementation of green chemistry practices will be severely hindered as only cursory guesses and assumptions can be made in such analyses. With missing data, the best that can be accomplished is to put upper bounds on RME values and lower bounds on E values.

The key set of parameters from the radial pentagon analysis for each reaction that are used as direct inputs in the second spreadsheet algorithm to assess the material efficiency of the entire synthesis plan are as follows: molecular weights of reagents and by-products, molecular weight of product, reaction yield, stoichiometric factor, total mass of excess reagents, mole scale of limiting reagent, mass of reaction solvent, and masses of all other auxiliaries. The use of the second spreadsheet algorithm is greatly facilitated by the construction of a synthesis tree for the plan according to the treatments described in [3, 5] that gives an exact count of the number of reaction stages, reaction steps, input materials, and branches in the plan. It is an excellent book-keeping and proofreading tool when reading research papers describing synthesis plans. A great advantage of using synthesis trees is that they give a visual representation of an entire plan in one simple diagram. Of particular note is the ease of analyzing convergent plans with multiple branches. Green metrics are determined from a synthesis tree diagram by reading it in the reverse sense from the target product node toward the input reagent nodes. In this way, the mole scales of all reagents are normalized relative to the target product, which is the basis scale for the entire plan. The method uses a simple connect-the-dots approach to determine the appropriate chain of reaction yields required to determine the mole scale of each reagent in a plan originating from any branch. It makes sense to follow the reverse approach since the target product node is the only one that is common to all branches in the tree diagram.

The algorithm for the computation of overall AE, RME, and E for any synthesis plan of any degree of complexity is given by a sequence of computations

beginning with the last step ($j = n$) and working toward the first step ($j = 1$) and is described elsewhere [22]. The overall E-factor may be partitioned into its three components as shown explicitly below:

$$E_{\text{total}} = E_{\text{kernel}} + E_{\text{excess}} + E_{\text{auxiliaries}} \quad (4)$$

where,

$$E_{\text{kernel}} = \frac{1}{p_n} \sum_j \left(\frac{1}{\prod_k^{n \rightarrow j} \varepsilon_k} \right) \left(\frac{p_j}{(\text{AE})_j} \right) [1 - \varepsilon_j(\text{AE})_j] \quad (5)$$

$$E_{\text{excess}} = \frac{1}{p_n} \sum_j \left(\frac{1}{\prod_k^{n \rightarrow j} \varepsilon_k} \right) \left(\frac{p_j}{(\text{AE})_j} \right) [(\text{SF})_j - 1] \quad (6)$$

$$E_{\text{auxiliaries}} = \frac{1}{p_n} \sum_j \left(\frac{1}{\prod_k^{n \rightarrow j} \varepsilon_k} \right) \left(\frac{c_j + s_j + \omega_j}{x_j^*} \right) \quad (7)$$

Symbol definitions:

x is number of moles of target product in synthesis plan, where x is set to 1 mole.

$\prod_k^{n \rightarrow j} \varepsilon_k$ is the multiplicative chain of reaction yields

connecting the target product node to the reactant nodes for step j as per synthesis tree diagram read from right to left (i.e., in the direction $n \rightarrow j$).

p_j is the molecular weight of product of step j .

ε_j is the reaction yield with respect to limiting reagent for step j .

$(\text{AE})_j$ is the atom economy for step j .

$(\text{SF})_j$ is the stoichiometric factor for step j that accounts for excess reagents used in that step.

$(c_j + s_j + \omega_j)$ is the sum of masses of auxiliary materials used in step j , namely the mass of catalyst, mass of reaction solvent, and mass of all other post-reaction materials used in the work-up and purification phases.

x_j^* is the experimental mole scale of limiting reagent in step j as reported in an experimental procedure.

The summation runs over the n reaction steps. A step begins with an isolated intermediate and ends with the following next isolated intermediate.

The relative magnitudes of these three components break down into the following order: $E_{\text{auxiliary}} \gg E_{\text{excess}} > E_{\text{kernel}}$. For synthesis plans, the hierarchy of contributor metrics that control overall material efficiency according to RME and E at the kernel molecular design stage is in descending order of the number of steps (n), reaction yields (ε), and atom economies (AE). When taking into account all materials used, the recovery and/or minimization of auxiliary materials in the work-up and purification phases plays the largest role in minimizing waste and, hence, the true overall magnitude of RME and E for a synthesis plan. Key assumptions implicit in the use of the algorithm described above are that an intermediate product collected in a given step is entirely committed as a reagent in the next step so that a true mass throughput can be assessed, and that reaction yield performances are invariant with mole scale, something that can only be checked by experiment but is a necessary assumption in implementing such an analysis. The results of synthesis plan algorithm are depicted graphically as a radial hexagon by direct analogy to the radial pentagon analysis. The key six parameters chosen are each fractions ranging from 0 to 1. For the material efficiency performance, there is overall AE, overall yield along the longest branch, and overall reaction mass efficiency. For the strategy efficiency performance, there is molecular weight fraction of reagents ending up as part of the structure of the final target product, the fraction of kernel waste arising from target bond reactions, and degree of convergence. The outer perimeter represents a perfectly green plan. The more the resultant actual hexagon is distorted toward the center, the worse is the plan. Again, strengths and weaknesses in a plan may be seen at once. In addition, the algorithm produces bar graphs that depict profiles of percentage of kernel and percentage of true waste distributions, atom economy, reaction yield, kernel reaction mass efficiency, cumulative mass of waste produced, hypsicity (oxidation level tracking), and target bond reactions per reaction stage. Bottlenecks in the plan may be spotted at once and then appropriate action taken to force optimization in the desired direction.

Strategy Metrics

Parameters that describe and track synthesis strategy for a given plan include: (1) target bond structure maps, (2) target bond forming reaction profiles from which the number of target bonds per reaction step is determined, (3) molecular weight fraction of sacrificial reagents used, and (4) hypsicity (oxidation level) index. A target bond structure map is simply a drawing of the target product structure showing which target bonds are made and at what reaction step. A target bond is denoted with a heavy connecting line, and its associated circled reaction step number is given alongside. From such a diagram, it is possible to trace the origin of each atom in the target structure back to the associated atoms in the reagents used. In effect, the set of reagent atoms is mapped onto the set of atoms comprising the target structure. From such a map, it is possible to determine the molecular weight fraction of sacrificial reagents whose atoms never get incorporated in the final target structure according to

$$f(\text{sac}) = 1 - \frac{\sum \text{MW}_{\text{reagents in whole or in part ending up in target product}}}{\sum \text{MW}_{\text{all reagents}}} \\ = 1 - \frac{(\text{AE})_{\text{overall}} \sum \text{MW}_{\text{reagents in whole or in part ending up in target product}}}{\text{MW}_{\text{target product}}}$$

Sacrificial reagents include those that serve as protecting groups, those that change the electronic states of key atoms so that skeletal building bond forming reactions are possible, those that are used to control stereochemistry, those that are used in substitution reactions to switch poor leaving groups into better ones, and those that are reducing or oxidizing agents that are used in *subtractive* redox reactions, that is, where oxygen atoms or hydrogen atoms are *removed* from a structure. The condition concerning redox reactions is important as it is *additive* redox reactions that are desirable to reduce $f(\text{sac})$ since they contribute oxygen or hydrogen atoms to the target structure. It is obvious that the ultimate goal is to minimize the magnitude of $f(\text{sac})$. One can deduce readily from the above equation that when the overall atom economy is unity, that is, when all atoms in all of the reagents used end up in the target structure, the two sums in the first part of the above equation become equal to each other, and hence, $f(\text{sac})$ is equal to zero. Also, from the target structure map, it is possible to construct a bar graph of the number of target bonds made versus the reaction

step count. Gaps in such bar graphs coincide with the use of sacrificial reagents in those steps and bars indicate productive steps. The number of target bonds made per step may be used as an indicator of synthetic efficiency and elegance since good synthesis strategies are characterized by fewer steps and the accomplishment of more target bonds made per step.

The target structure map also provides the set of atoms in the target structure that are involved in bond-making and bond-breaking processes throughout the synthesis. These atoms are precisely the ones connected by the heavy lines in the target structure map. It is possible to trace the oxidation numbers of these atoms from the target structure back to the progenitor reagent used via the intervening intermediate products. In a manner similar to the determination of the molecular weight first moment building up parameter, the difference between the oxidation number of an atom in an intermediate structure and that same atom in the final target product is determined for each atom involved in this special set of atoms as a function of reaction stage. This idea of tracking the changes in oxidation state, or hypsicity, (Gk: *hypsos*, meaning level or height) of key atoms involved in bond-making and bond-breaking steps was introduced by Hendrickson [30]. He proposed that good synthesis plans aim for the *isohypsic* condition, which is characterized by a zero net change in oxidation state of all atoms of starting materials and intermediates involved until the target product is reached. This can be achieved by designing synthesis plans that eliminate redox reactions entirely as it is consistent with the conclusion that such reactions are to be minimized in a plan because they are the most material inefficient class of chemical reaction and therefore contribute to significant attenuations in kernel and global RMEs. If these cannot be avoided due to practical considerations, then the next best thing to achieve the *isohypsic* condition is to strategically sequence redox reactions in such a way that for every increase in oxidation level of an atom occurring in a step, it is matched by a concomitant decrease in oxidation level of equal magnitude in the next step, or vice versa. This cuts down on the accumulation of excess gains or losses in oxidation level of atoms, as the case may be, in starting materials and intermediates with respect to the oxidation levels of those atoms in the final target molecule over the course

of the synthesis. This second option of achieving the *isohypsic* condition is purely due to an algebraic cancellation of all the “ups” and “downs” in oxidation state changes and, hence, will coincide with appreciably higher kernel E-factor values and lower kernel RME values for such plans since generally redox reactions are least material efficient. The first option, however, will coincide with lower kernel E-factor values and higher kernel RME values since redox reactions would be completely eliminated.

Formally, we may define a hypsicity index, HI, as

$$HI = \frac{\sum_{\text{stages},j} \left[\sum_{\text{atoms},i} \left[(\text{Ox})_{\text{stage},j}^{\text{atom},i} - (\text{Ox})_{\text{stage},N}^{\text{atom},i} \right] \right]}{N + 1} = \frac{\sum_{\text{stages},j} \Delta_j}{N + 1}$$

where (Ox) represents the relevant oxidation number of an atom. If HI is zero, then the synthesis is *isohypsic*. If HI is positive valued, then to get to the target molecule a net reduction is required over the course of the synthesis since an accumulated gain in oxidation level has resulted. Such a condition is termed *hyperhypsic*, by analogy with the term *hyperchromic*, which describes increases in intensities of absorption bands in spectroscopy. Conversely, if HI is negative valued, then to get to the target molecule a net oxidation is required over the course of the synthesis since an accumulated loss in oxidation level has resulted. Such a condition is termed *hypohypsic*, again by analogy with the term *hypochromic*. It is important to note that changes in oxidation number can occur for atoms in reactions that are not formally classified as reductions or oxidations with respect to the substrate of interest. A good example of this is the Grignard reaction which is classified as a carbon–carbon bond forming reaction and yet involves a formal oxidation with respect to magnesium in the preparation of the Grignard reagent. Another is electrophilic aromatic substitution, which begins with an oxidation state of -1 for the ArC-H carbon atom, which then increases to $+1$ when hydrogen is substituted for chlorine, for example. The hypsicity index therefore accounts for all such changes in oxidation numbers of atoms regardless of the reaction type.

The following sequence of steps may be followed to determine HI for a synthesis plan:

1. Enumerate atoms in the target structure that are only involved in the building up process from

corresponding starting materials according to the structure map. This set of atoms defines those that are involved in bonding changes occurring in the relevant reaction steps.

2. Work backwards intermediate by intermediate to trace the oxidation numbers of the above set of atoms back to original starting materials as appropriate following the reaction stages back to the zeroth stage.
3. For each key atom, i , in each reaction stage, j , determine the difference in oxidation number of that atom with respect to what it is in the final target structure. Hence,

$$(\text{Ox})_{\text{stage},j}^{\text{atom},i} - (\text{Ox})_{\text{stage},N}^{\text{atom},i}$$

4. Sum the differences determined in step (3) over all key atoms in stage j . This yields the term
$$\sum_{\text{atoms},i} \left[(\text{Ox})_{\text{stage},j}^{\text{atom},i} - (\text{Ox})_{\text{stage},N}^{\text{atom},i} \right] = \Delta_j.$$
5. Finally, take the sum $\sum_{\text{stages},j} \Delta_j$ over the number of stages and divide by $N + 1$ accounting for the extra zeroth reaction stage.

From the above discussion, for a material efficient synthesis plan, it is not a sufficient condition to just aim for an HI value of zero. What matters is that as many of the atoms as possible in oxidizing and reducing agents end up in the target structure. These can only arise from oxidation and reduction reactions that are of the *additive* type where oxygen and hydrogen atoms are added to a structure in a reaction and remain there until the final target structure is reached, and not of the *subtractive* types as described earlier in the discussion of sacrificial reagents. So, it is possible to have HI values that are strongly positive or negative and still be material efficient so long as those key atoms are incorporated as part of the final target structure. It is preferable to have a hysicity profile that exhibits either a steadily increasing or a steadily decreasing oxidation level rather than an undulating one. Increases and decreases in oxidation levels parallel the types of redox reactions employed in a plan. This will of course depend on the oxidation levels of atoms in the selected reagents used throughout the synthesis. Hence, careful correlation of HI values and overall AE, kernel RME, and $f(\text{sac})$ values need to be made to understand the synergy between

material efficiency and oxidation level changes. It is impossible to make inferences solely on the basis of the magnitude of the HI. One needs to examine the shape or distribution of the bar graph that tracks the changes in oxidation level as function of reaction stage for a given synthesis plan.

An Example: Sertaline Plans

Seven synthesis plans for sertraline were examined – four industrial (Pfizer G1, Pfizer G2, Pfizer G3, and Gedeon Richter) and three academic (Buchwald, Lautens, and Zhao). Table 1 summarizes the material efficiency metric performances, Table 2 summarizes the E-factor breakdown, and Table 3 summarizes the strategy efficiency metric performances. From these data, it appears that the overall best material efficient synthesis plan is that of Zhao since it has the highest overall yield, atom economy, and overall kernel RME. Consequently, it produces the least kernel and overall masses of waste. It also uses the least number of input materials. The Pfizer plans show a nice progression from G1 to G3 with a sevenfold decrease in overall E-factor from G1 to G2 and a further 2.5-fold decrease from G2 to G3. The Zhao, Pfizer G3, Gedeon Richter, and Buchwald have the least number of steps at five. The Buchwald and Pfizer G3 plans have all of their kernel waste coming from target bond forming reactions; whereas, the Lautens plan has most of its kernel waste coming from sacrificial reactions. The Lautens plan also has the worst overall atom economy. In terms of E-factor contributors, the Zhao plan has the least E-auxiliary whereas the Pfizer G3 plan has the least E_{kernel} and the Gedeon Richter plan has the least E_{excess} contributions. The overall most wasteful Pfizer G1 plan has the highest contributors in all three categories. From a strategy perspective, it is less clear-cut to decide the overall best plan. The Pfizer G3, Buchwald, and Zhao plans show the greatest degree of building up from starting materials toward product throughout the course of their syntheses. The Gedeon Richter plan has the least degree of asymmetry. The Pfizer G3 plan has the highest degree of convergence. The Pfizer G3 plan has highest number of target bonds made per reaction stage. The Pfizer G3 and Buchwald plans have all reaction stages producing at least one target bond; whereas, almost half the reaction stages in the Lautens plan do

Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Table 1 Summary of material efficiency metrics for the seven plans to produce sertraline ranked in descending order according to percentage of kernel reaction mass efficiency

Plan	Type	N ^a	n ^b	I ^c	f ^{*d}	Overall yield (%)	AE (%)	Kernel RME (%)
Zhao ^e	linear	5	5	9	0.541	35.0	43.0	18.1
Pfizer G3 ^f	linear	5	5	14	1	25.0	39.0	11.2
Gedeon Richter ^f	linear	5	5	11	0.639	20.5	36.6	9.4
Buchwald ^e	linear	5	5	12	1	19.2	38.9	8.6
Pfizer G2 ^f	linear	8	8	17	0.894	8.7	34.9	4.3
Lautens ^e	convergent	11	13	28	0.389	14.1	7.1	1.9
Pfizer G1 ^f	linear	10	10	18	0.731	1.8	26.7	1

^aNumber of reaction stages

^bNumber of reaction steps

^cNumber of input reagents

^dFractional kernel waste contribution from target bond forming reactions

^eTarget is (+)-sertraline

^fTarget is (+)-sertraline hydrochloride

Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Table 2 Summary of mass of waste and E-factor breakdown for seven plans to produce sertraline ranked in ascending order^a

Plan	E _{kernel}	E _{excess}	E _{auxiliary}	E _{total}	Kernel mass of waste (kg)	Total mass of waste (kg)
Zhao ^b	4.54	89.21	64.64	158.38	1.4	48.4
Pfizer G3 ^c	7.92	55.21	226.23	289.76	2.7	99.2
Gedeon Richter ^c	9.59	42.07	339.53	391.19	3.3	133.9
Buchwald ^b	10.56	90.25	352.24	453.05	3.2	138.6
Pfizer G2 ^c	22.34	163.56	539.67	725.56	7.6	248.4
Lautens ^b	51.99	114.58	2424.98	2591.55	15.9	792.8
Pfizer G1 ^c	103.95	350.30	4536.58	4990.82	35.6	1708.6

^aBasis is 1 mole target product

^bTarget is (+)-sertraline

^cTarget is (+)-sertraline hydrochloride

not produce target bonds. The Zhao and Gedeon Richter plans have the least number of oxidation changes throughout the course of their synthesis plans, whereas, the Lautens plan has the most.

Figures 2 and 3 show the synthesis schemes for the best overall performing Zhao and Pfizer G3 plans. The Zhao plan involves a Friedel–Crafts alkylation in step 1 to give a racemic product, an asymmetric ketone

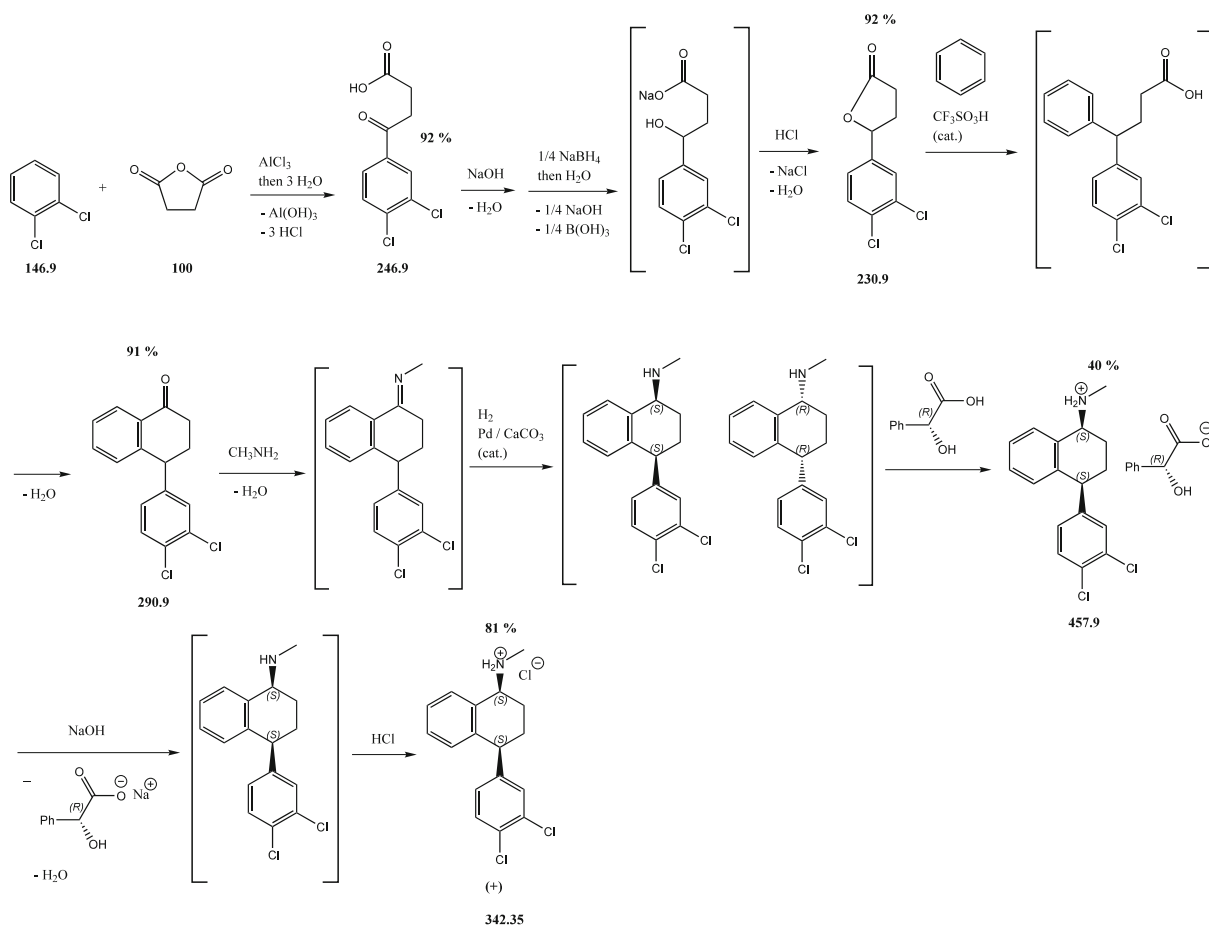
reduction in step 2 with a chiral proline organocatalyst to give an optically active trans-tetralol intermediate, alcohol oxidation in step 3, imination in step 4, and catalytic hydrogenation in step 5. The Pfizer G3 plan involves a Friedel–Crafts acylation in step 1, a tandem reduction-lactonization sequence in step 2, a tandem lactone ring opening Friedel–Crafts acylation in step 3 to create the 1-tetralone ring system, a telescoped

Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Table 3 Summary of strategy efficiency metrics for the seven plans to produce sertraline

Plan	$\mu 1$ (g/mol)	β	δ	f(sac)	HI	B/M	f(nb)	UD
Zhao ^a	−60.98	0.738	0.435	0.495	+1	1.20	0.33	8
Pfizer G3 ^b	−96.77	0.781	0.508	0.540	+2	1.60	0	12
Gedeon Richter ^b	−58.69	0.734	0.476	0.521	+1.33	1.40	0.20	8
Buchwald ^a	−85.98	0.760	0.486	0.358	+2.33	1.40	0	12
Pfizer G2 ^b	−72.83	0.839	0.426	0.562	+1.56	1.00	0.25	12
Lautens ^a	−15.81	0.776	0.471	0.764	+1	0.92	0.45	24
Pfizer G1 ^b	−50.04	0.848	0.389	0.579	+1.64	0.90	0.30	10

^aTarget is (+)-sertraline

^bTarget is (+)-sertraline hydrochloride



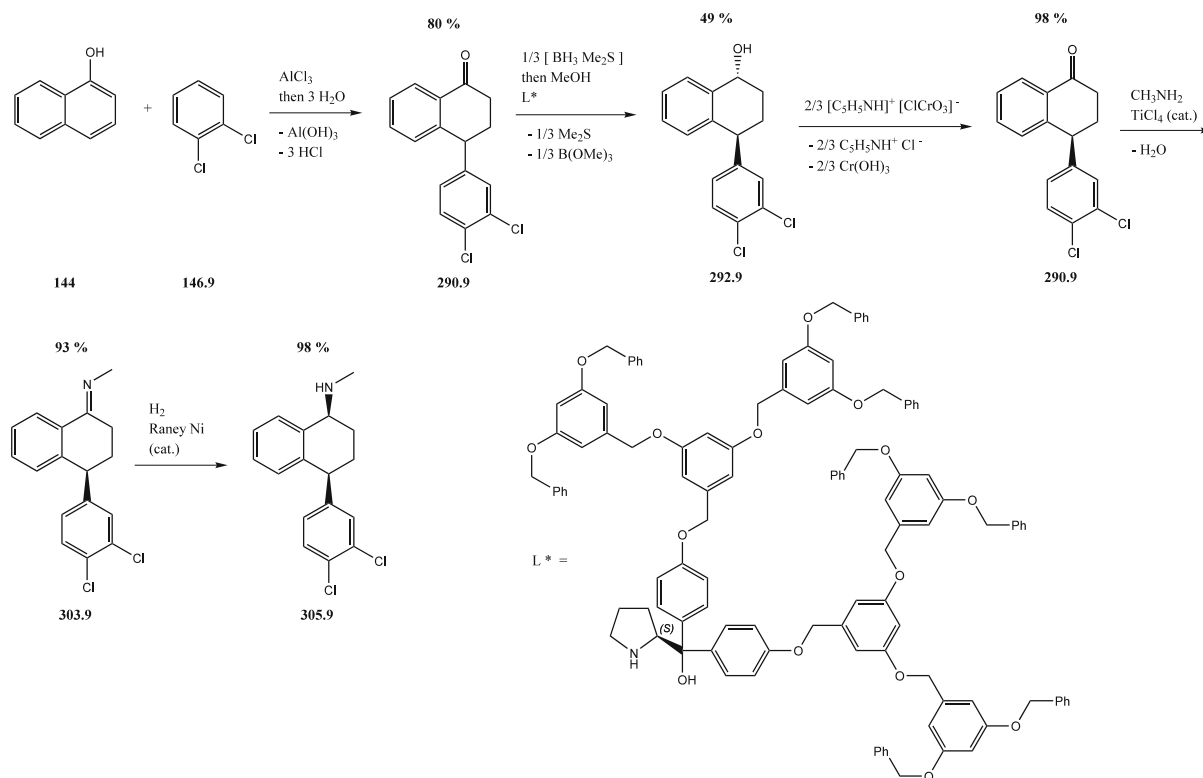
Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 2 Pfizer G3 synthesis plan

imation-hydrogenation-diastereomeric salt resolution sequence in step 4, and, finally, isolation of the hydrochloride salt of the correct *cis*-amine product in step 5. The corresponding synthesis tree diagrams are shown in Fig. 4. Both plans are linear.

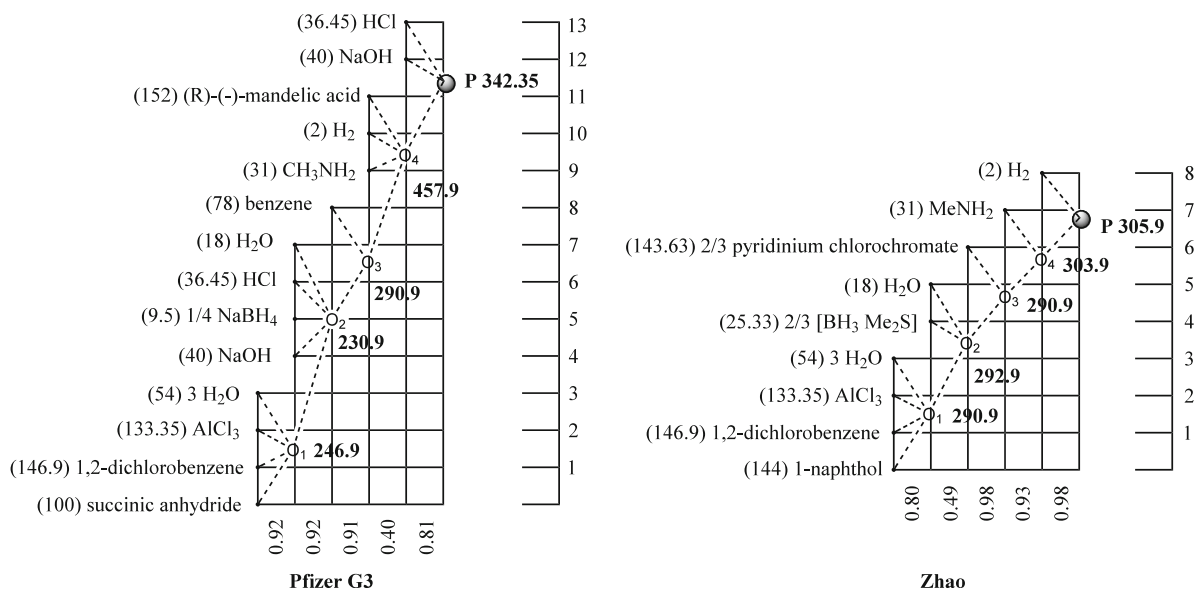
Figures 5 and 6 show the radial pentagons for each reaction in the Zhao and Pfizer G3 plans. For the Zhao plan, the bottlenecks are the low atom economy in step 1, the low yield in step 2, the excess reagent usage in steps 1, 2, and 4, and the high auxiliary material usage in steps 1, 3, and 4. For the Pfizer G3 plan, the bottlenecks are the low atom economy in step 1, the low yield in step 4, the high excess reagent usage in steps 1 to 3, and the high auxiliary material consumption in steps 3 to 5.

The numerical data presented in Tables 1–3 are conveniently displayed graphically in Fig. 7, which shows the radial hexagons for all seven plans. It is easily observed that the overall best performing Zhao plan

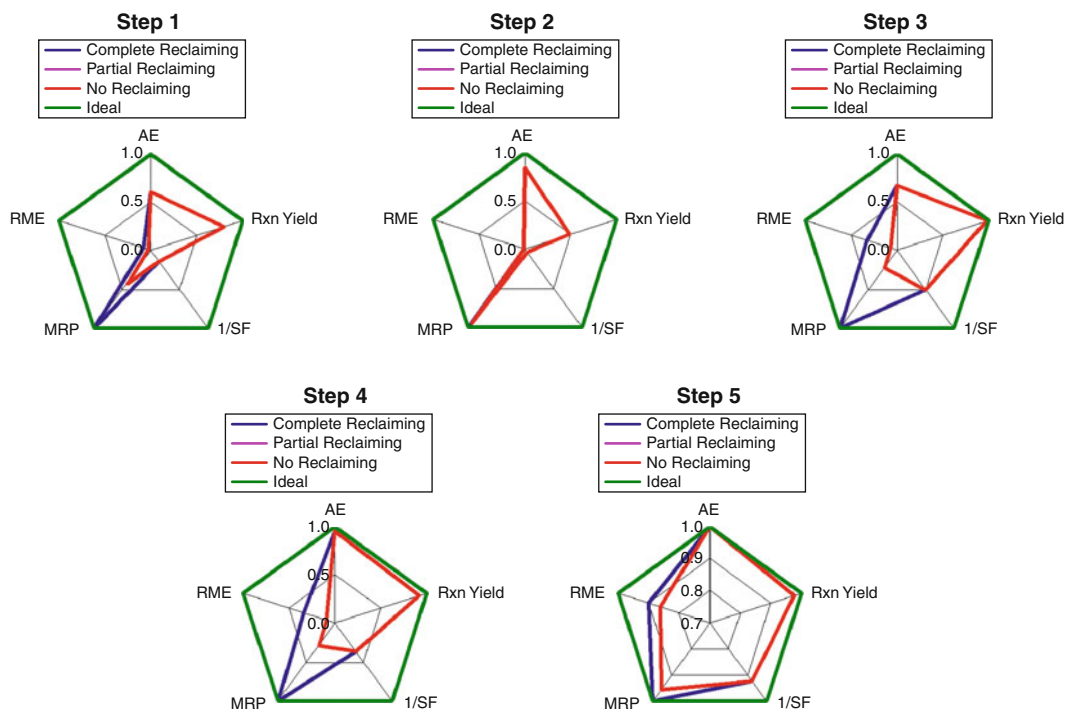
has the least distorted hexagon with the largest area compared to the other plans. The Lautens plan has the least area. Figures 8 and 9 show the kernel and total waste distribution profiles for each plan. The kernel waste distributions account for wastes coming from reaction by-products, side products, unreacted starting materials, and excess reagents. The total waste distributions add on the contributions from auxiliary materials. Reaction steps producing significant amounts of waste from auxiliary materials may be directly correlated with their corresponding MRP performances from their radial pentagons. On comparing both kinds of distributions for each plan, it is observed that the Zhao plan is the only one that has consistently the same shape. Every synthesis plan is made up of reactions that either make target bonds found in the final product or not. In terms of optimization, the goal is to minimize both the global waste and the proportion arising from sacrificial nonproductive reactions at the



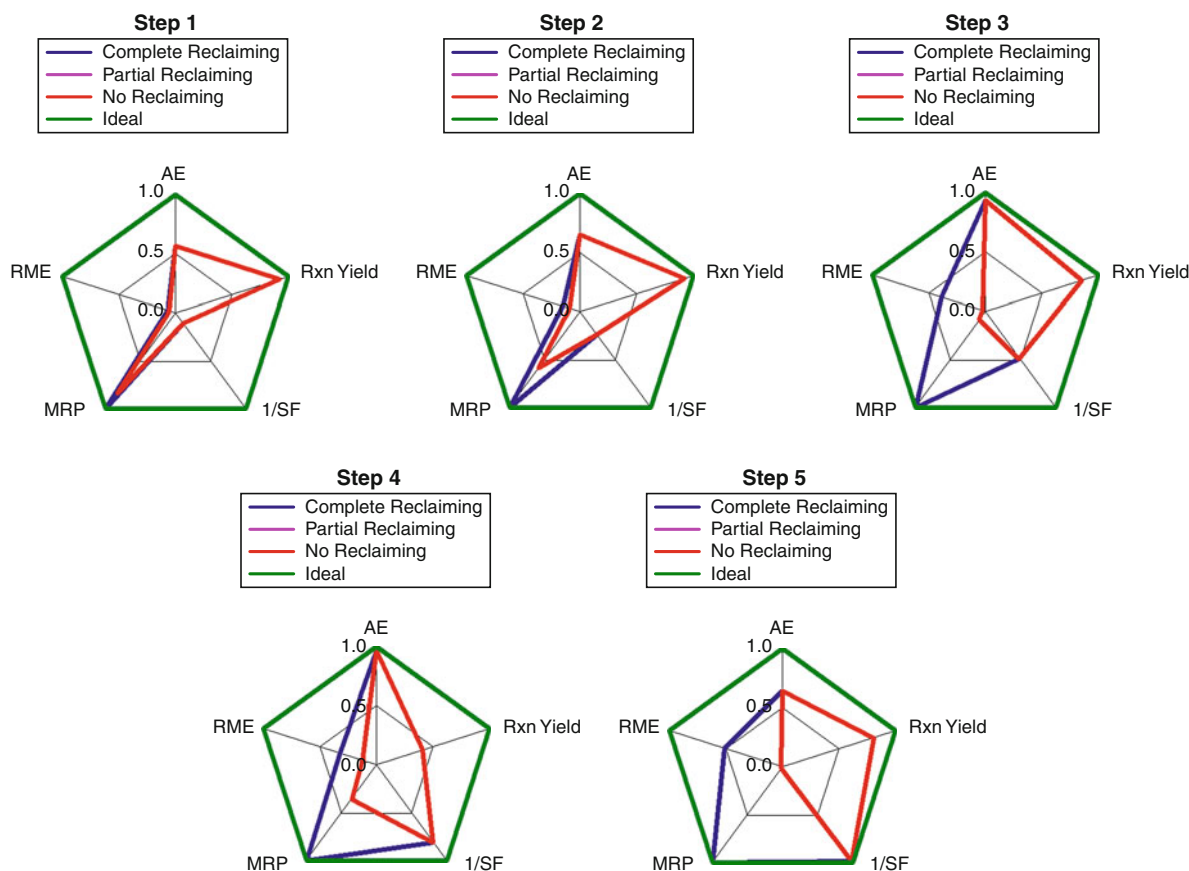
Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 3
Zhao synthesis plan



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 4
Synthesis trees for Pfizer G3 and Zhao plans



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 5
Radial pentagons for Zhao plan



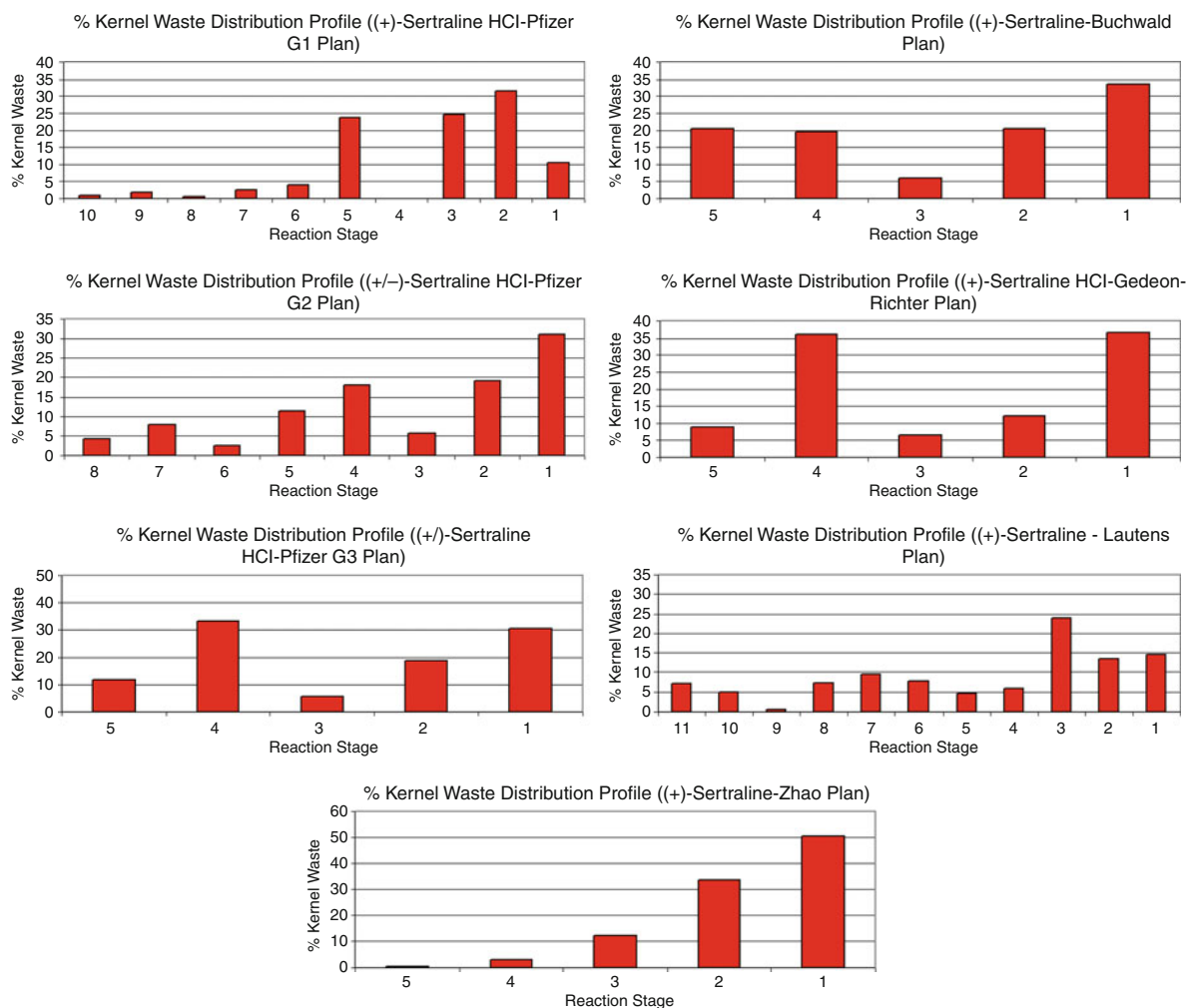
Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 6
Radial pentagons for Pfizer G3 plan

same time. A disastrous situation to avoid is coming up with a synthesis plan that produces a lot of waste, most of which originating from sacrificial reactions such as redox adjustments and excessive use of protecting groups. Figure 10 shows the kernel mass of waste profile for all sertaline plans as a function of the global waste produced and the waste contributions from target bond forming and sacrificial reactions. The lengths of the bars correlate directly with the overall kernel E-factors and each bar, in turn, is subdivided into these two broad groups of reactions. The molecular weight building up profiles are given in Fig. 11. It is easily observed that the Buchwald and Zhao plans exhibit a high degree of building up since no intermediate along their reaction sequences has a molecular weight exceeding that of the target product sertaline. From the hypsicity profiles shown in Fig. 12, it is

observed that the Buchwald, Pfizer G2, and Pfizer G3 are hyperhypsic with steadily decreasing oxidation steps which is consistent with the additive reduction reactions employed in their plans. It is obvious that the Lautens plan has the most oxidation number changes compared to the other plans. The target bond profiles and maps for all plans are shown in Figs. 13 and 14. These profiles confirm the conciseness and, therefore, the strategic efficiencies of the Buchwald and Pfizer G3 plans since there are no gaps in their respective bar graphs. From Fig. 14, it is observed that the plans break down into four distinct strategies in constructing the 1,2,3,4-tetrahydro-naphthalene ring system: (1) the Pfizer G2, Pfizer G3, and Buchwald plans involve [4 + 2] cycloadditions via Friedel–Crafts acylations; (2) the Gedeon Richter and Zhao plans begin with 1-naphthol as starting material so the ring frame is



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 7
Radial hexagons for seven synthesis plans of sertraline

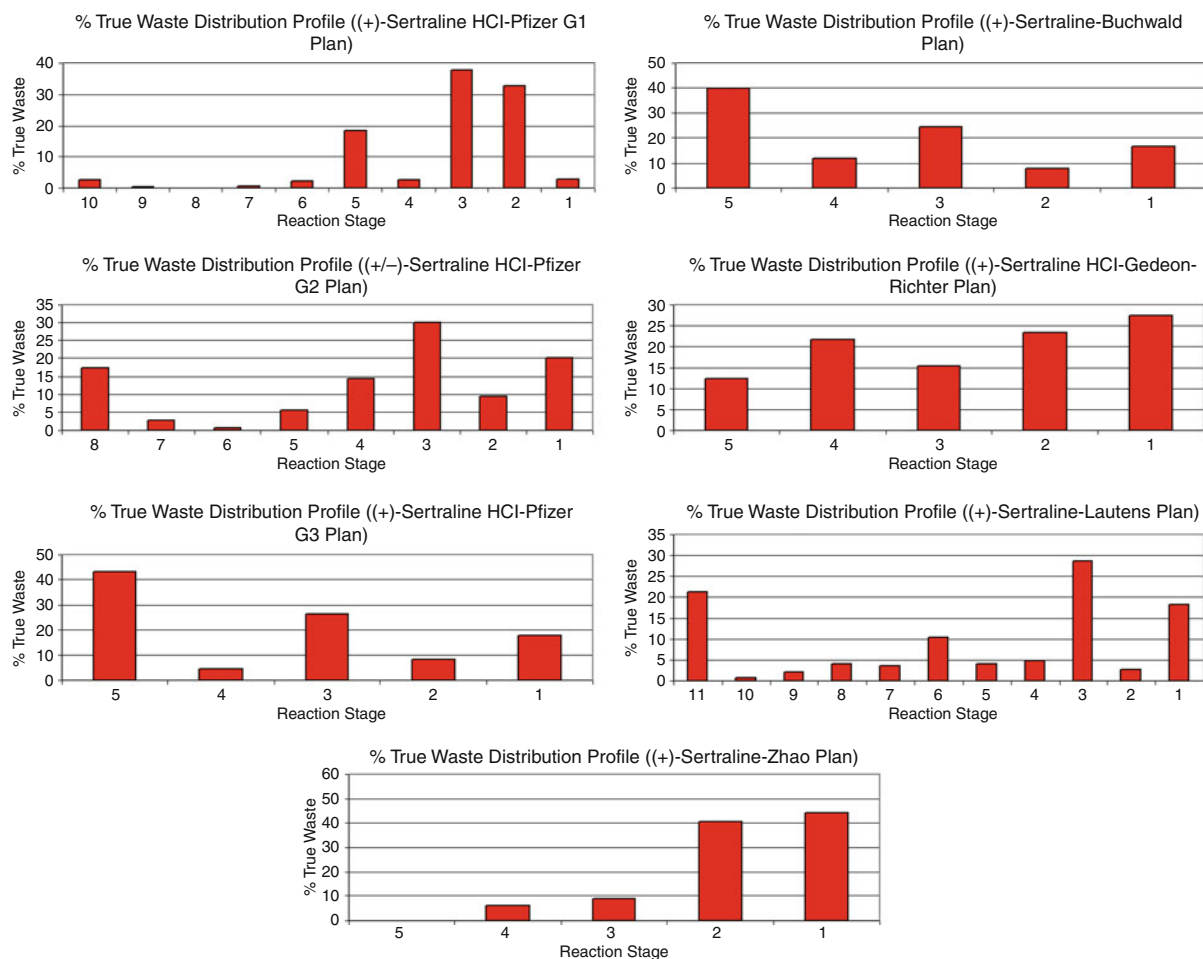


Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 8

Kernel waste distribution profiles for synthesis plans of sertraline

already set; (3) the Pfizer G1 plan involves building up the saturated ring by chain extension in step 2 via an aldol condensation and then a [6 + 0] ring closure in step 5 via an intramolecular Friedel–Crafts acylation; and (4) the Lautens plan involves a unique ring-constructing sequence via a Diels–Alder [2.2.1] adduct, which then undergoes ring opening via an asymmetric reduction as shown in Fig. 15. Though the Lautens strategy clearly demonstrates ingenuity and novelty, it

unfortunately failed to translate this attribute into an overall material efficient plan that could compete with the other industrial plans beginning from the third-generation feedstock 1-naphthol available in two steps from the coal tar product naphthalene. The main reasons for this are the high number of reaction steps (11 versus 5) and the high proportion of sacrificial nontarget bond forming reactions involved (see Table 1). This ugly mismatch between demonstration



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 9
Total waste distribution profiles for synthesis plans of sertraline

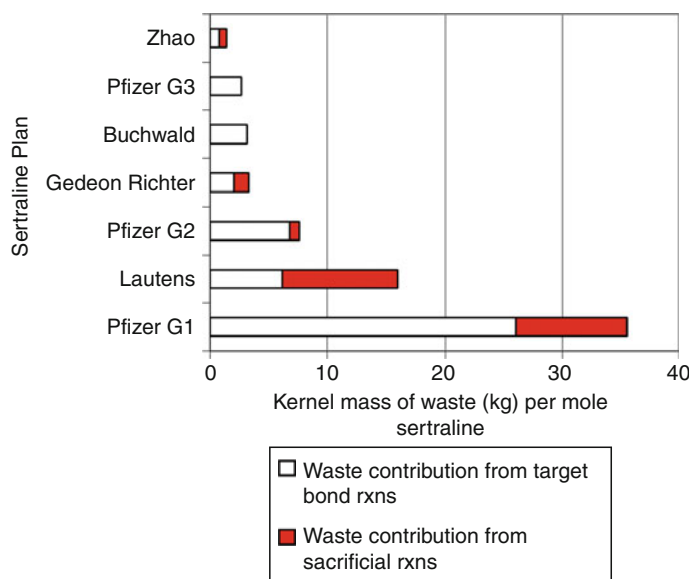
of novelty in a few steps and overall low-material-efficiency performance in a plan occurs unfortunately all too often. Hence, the achievement of an optimum synergy between the two is indeed hard to achieve within the context of realizing a truly global “green” and hence cost-effective synthesis plan. However, fruitful opportunities for making new reaction discoveries and linking existing reactions in new combinations always exist and will help to make this goal a reality as synthetic organic chemistry continues to expand (Fig. 16).

Summary of Optimization

Characteristics of a “good” synthesis plan that merge both optimum material and strategy efficiencies are given below.

Material Efficiency

1. Minimize overall waste by first identifying major waste contributors such as solvent, work-up, and purification materials used and then minimizing them where possible.
2. Waste contribution from target bond forming reactions should exceed that from sacrificial reactions.
3. Maximize overall reaction yield along longest branch.
4. Maximize overall reaction mass efficiency so that kernel RME $\geq 60\%$ for each reaction in a synthesis plan.
5. Maintain SF as close to one as far as possible to reduce impact of excess reagent consumption.



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 10

Kernel mass of waste profile for synthesis plans showing contributions from target bond reactions and sacrificial reactions

Strategy Efficiency

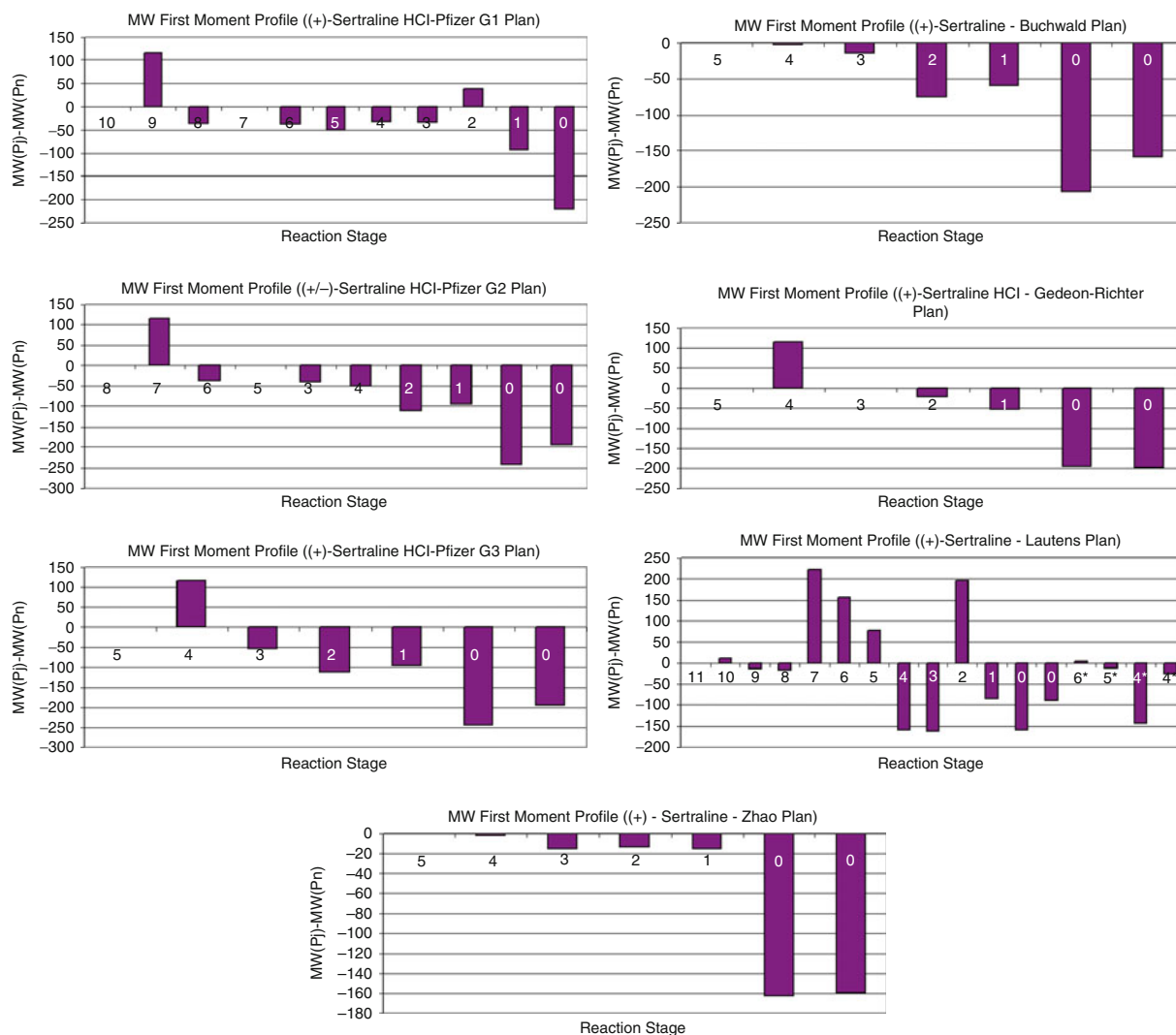
1. Number of reaction steps that form target bonds should exceed number of reaction steps that are sacrificial reactions.
2. Minimize overall number of reaction steps.
3. Maximize overall atom economy so that $AE \geq 60\%$ for each reaction in a synthesis plan.
4. Maximize degree of convergence.
5. Maximize the number of direct synthesis-type reactions such as carbon-carbon and non-carbon-carbon bond forming reactions, additive redox reactions, multicomponent, tandem, domino, or cascade reactions, strategic rearrangements particularly with respect to ring constructions, and non-sacrificial substitution reactions.
6. Minimize the number of indirect synthesis-type reactions such as eliminations or fragmentations, subtractive redox reactions, and nonstrategic rearrangements.
7. Minimize fractional contribution of sacrificial reagents.
8. Minimize degree of asymmetry.
9. Design plans with large negative molecular weight first moments so that plans begin from low

molecular weight starting materials and progressively build up toward the molecular weight of the target molecule with minimal overshoots above the molecular weight of the target product.

10. With respect to redox economy: aim for $HI = 0$ by eliminating all redox reactions, or aim for $HI > 0$ with a steadily decreasing hypsicity profile using additive-type reduction reactions, or aim for $HI < 0$ with a steadily increasing hypsicity profile using additive-type oxidation reactions.

Thresholds and Probabilities

From the connecting relationships $RME = \frac{1}{1+E}$ and $AE = \frac{1}{1+E_{mw}}$, it is possible to set a threshold for “greenness” for individual reactions at the kernel level (excluding all auxiliary materials) as $AE \geq 0.618$ so that $AE \geq E_{mw}$, and $RME \geq 0.618$ so that $RME \geq E$, respectively. This threshold which exactly corresponds to the golden ratio equal to $\phi = \frac{\sqrt{5}-1}{2}$ is found by equating RME and E in the former expression and AE and E_{mw} in the latter. This is the justification for putting a 60% minimum cutoff for both AE and RME in the optimization

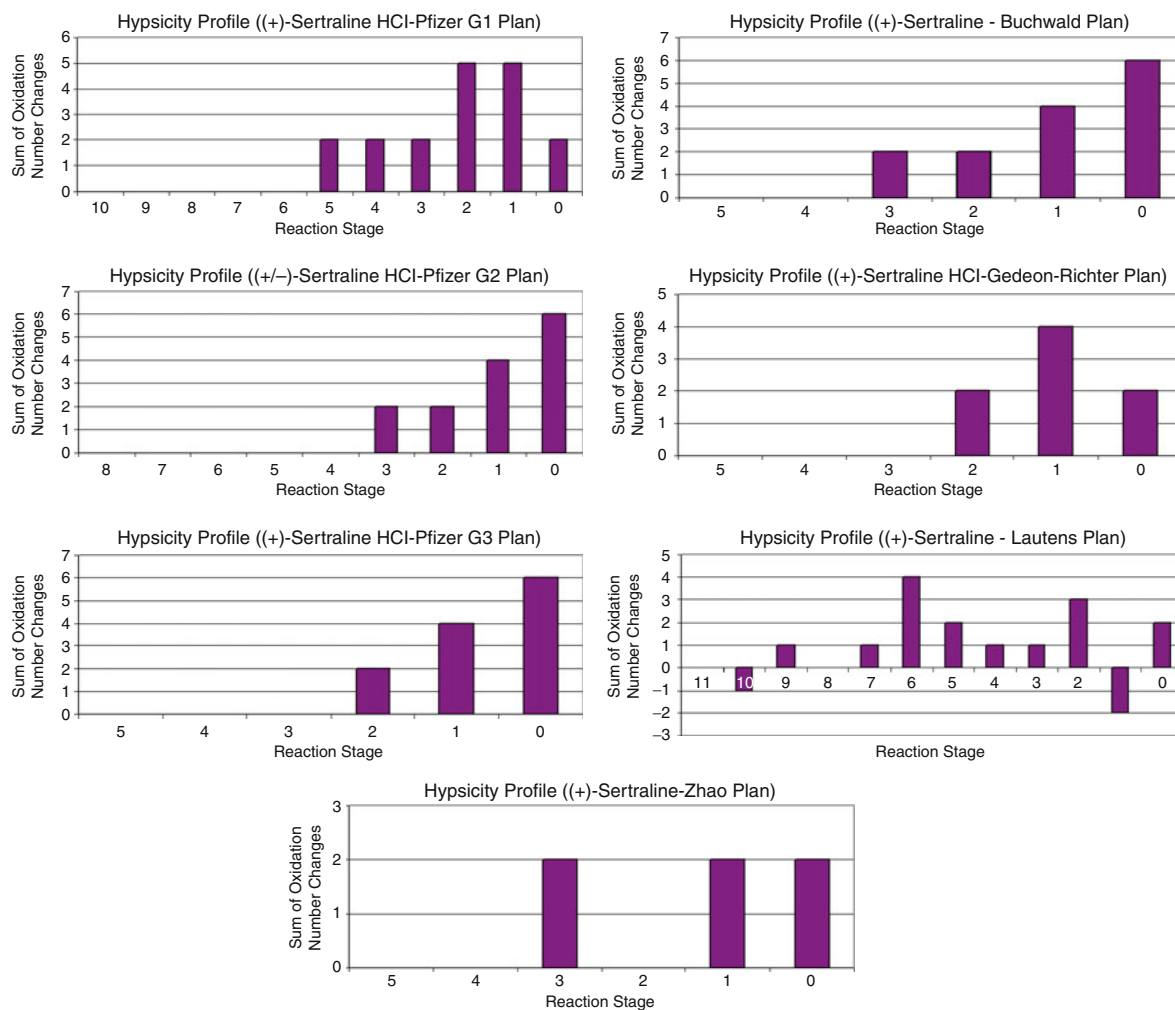


Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 11
Molecular weight first moment profiles for synthesis plans of sertraline

recommendations given in section 6. The implication of this from a marketing point of view is that aiming for “green” chemistry is really aiming for “golden” chemistry. If individual reactions fail to meet this criterion at the kernel level, then there is no hope that they will meet it when all auxiliary materials are taken into account.

For an individual reaction, the kernel RME is given by $RME = (\varepsilon)(AE)$, which is a reduced form of Eq. 1 when $SF = 1$ (no excess reagents) and $MRP = 1$

(all auxiliaries recovered or eliminated). If a threshold value of ϕ is set for the kernel RME, then it is possible to define the probability of achieving this target RME given a reaction’s AE and range of possible reaction yield performances. For any given chemical reaction, the magnitude of AE is fixed while the reaction yield is in principle variable over the full range 0–100%. Applying the inequality $RME = (\varepsilon)(AE) \geq 0.618$ implies that $(AE) \geq \frac{0.618}{\varepsilon}$. Figure 17 shows plots of AE versus reaction yield under two scenarios, depending on

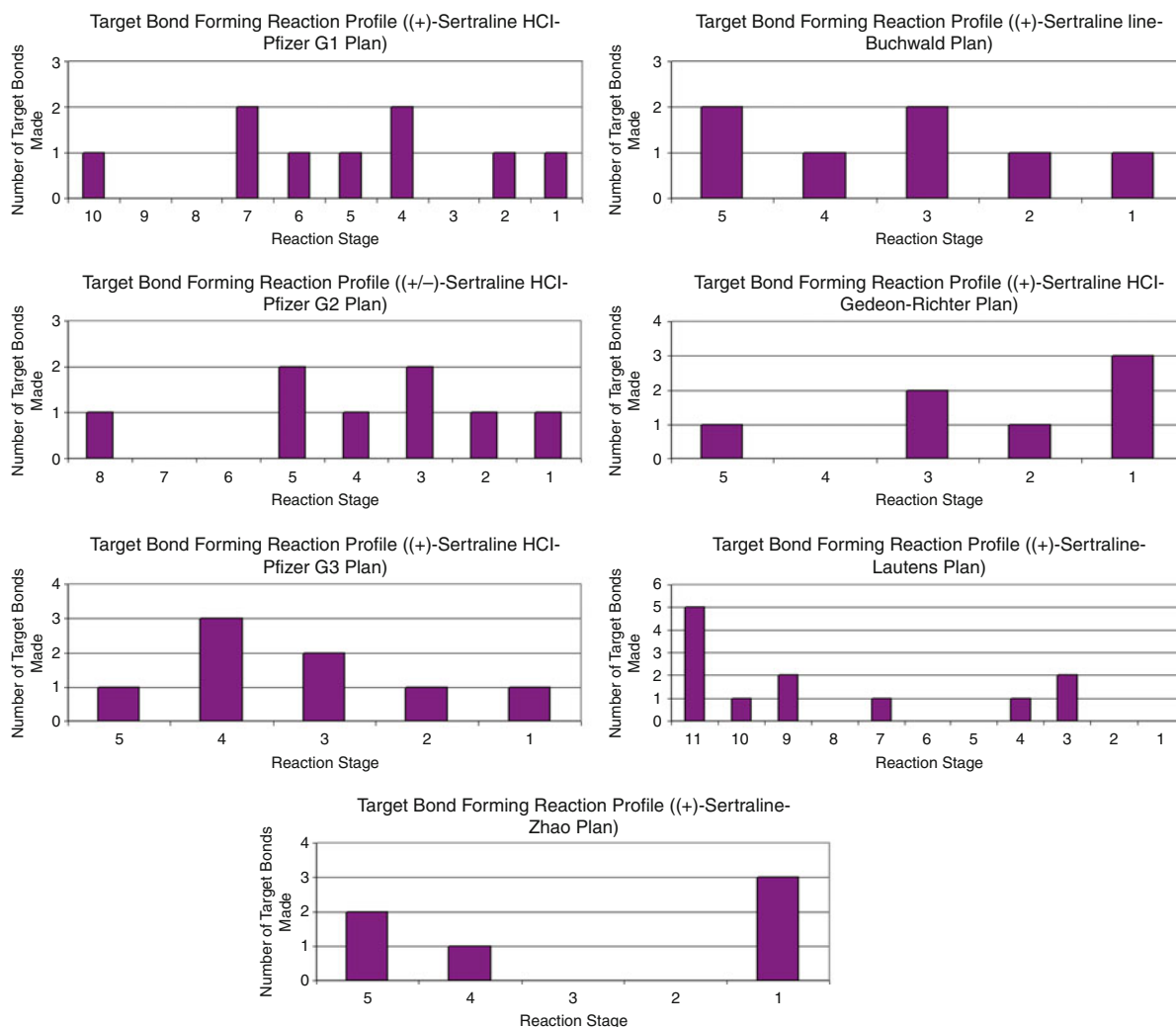


Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 12

Hypsicity profiles for synthesis plans of sertraline

whether a reaction's atom economy, $(AE)^*$, is below or above the threshold value of 0.618. The curved line represents the equation $AE = \frac{0.618}{\epsilon}$. The region above this line satisfies the condition $RME > 0.618$, and the region below the line satisfies the condition $RME < 0.618$. If a reaction's AE is below 0.618, then the probability of it achieving an RME above 0.618 is zero no matter how high the reaction yield is. This is obvious because the horizontal line does not intersect the curve $AE = \frac{0.618}{\epsilon}$. However, if a reaction's AE is above 0.618, then the probability of it achieving this goal is given by the length of the bolded line segment, or $p = 1 - \frac{0.618}{(AE)^*}$.

So, for example, if a reaction has $(AE)^* = 0.8$, then $p = 0.23$ or 23%. The best possible scenario would yield a probability of only $p = 38\%$ for a reaction with 100% AE. This low probability may be amplified considerably if we consider a minimum cutoff for the reaction yield as well that is better than 0%. This means case II shown in Fig. 17 may be further subdivided into two scenarios given by the conditions $0 < \epsilon_{\min} < \frac{0.618}{(AE)^*}$ and $\frac{0.618}{(AE)^*} < \epsilon_{\min} \leq 1$ as shown in Fig. 18. In the former case, $p = \left(1 - \frac{0.618}{(AE)^*}\right) / (1 - \epsilon_{\min})$ and in the latter, $p = 1$. So, if a reaction proceeds with a minimum yield of 70% and it has an AE of 80%, then the probability of



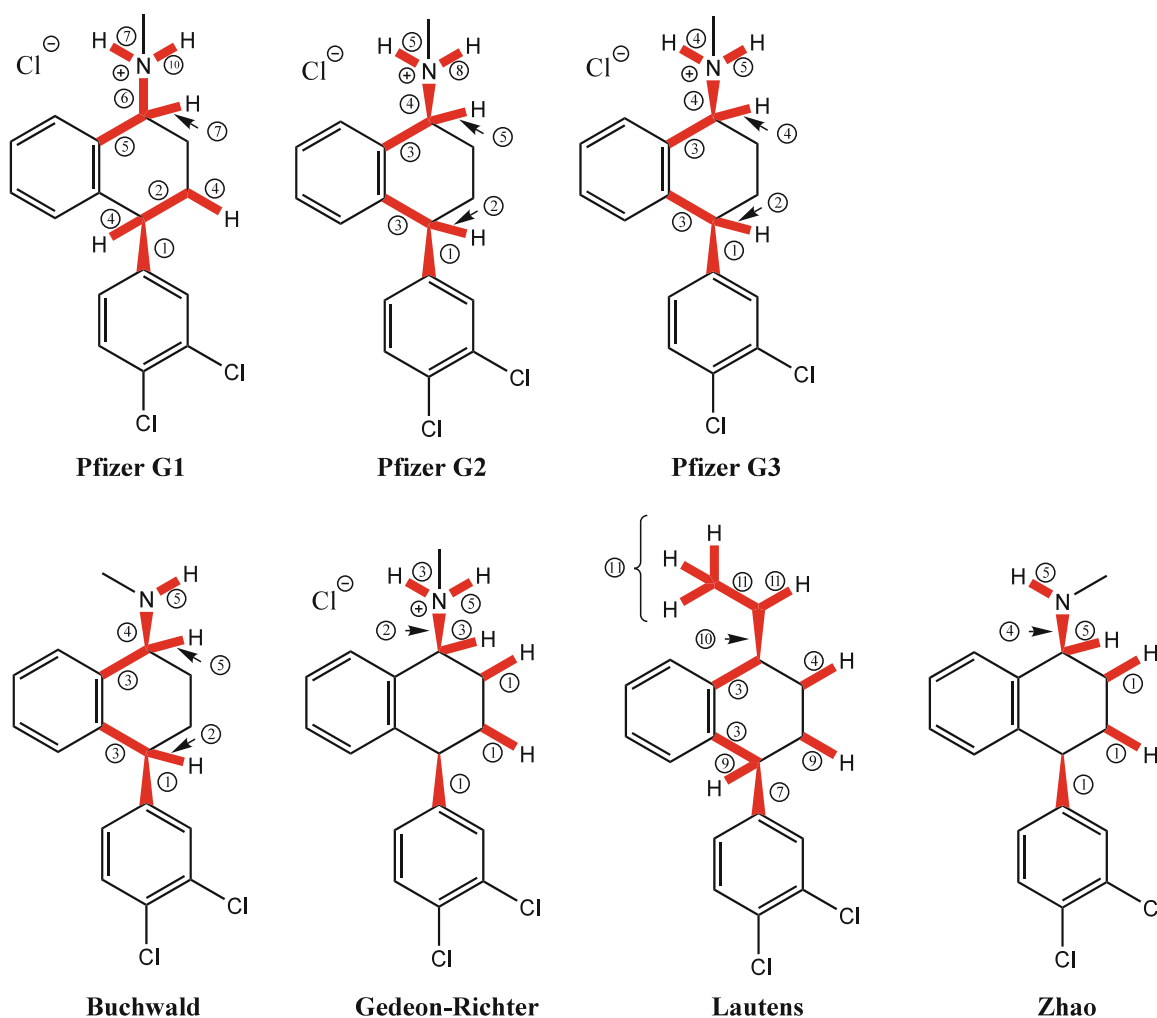
Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 13
Target bond profiles for synthesis plans of sertraline

achieving an RME of 0.618 improves over threefold from $p = 0.23$ to $p = 0.76$. If the minimum yield is no worse than $0.618/0.8 = 0.77$ (or 77%), then it is certain that the goal condition is attainable. These simple calculations clearly illustrate the importance of achieving both atom economies and reaction yields as high as possible in order to increase the chances of achieving even a modest target threshold RME of 61.8%. In a nutshell, synthesis plans composed of reactions with high atom economies and high yields working together have the best chance of achieving high kernel reaction mass efficiencies.

Future Directions

The future success of applying green chemistry thinking and metrics to synthesis design and optimization will depend on:

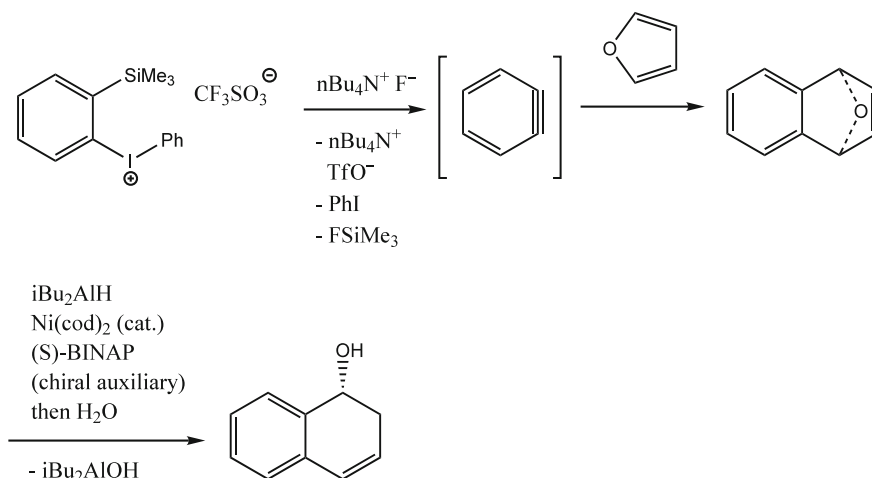
1. Making use of multicomponent and tandem one-pot reactions as central theme reactions in synthesis planning
2. Disclosing all material and energy consumption parameters for each reaction in a given plan (especially masses of chromatographic solvents, drying agents, work-up wash solutions, and reactant



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 14

Target bond maps for synthesis plans of sertraline

- gases) and clamping down on careless and intentional errors in publications, particularly experimental sections and supplementary materials of journal articles, with respect to mass and mole amounts of materials used and reaction yields so that true assessments of RME and E-factors can be made
- Implementing the radial pentagon spreadsheet algorithm as a standard proofreading tool for reviewers of synthesis papers to trap errors described in (2) and help authors in improving the standard and reliability of their reported procedures
- Mandating the reporting of green metrics as proof of greenness and part of the standard protocol in reports of synthesis plans in the literature, especially if plans are advertised as “green-er” than prior published plans and hence reduce false claims of “greenness”
- Accepting that optimization is a continuous ongoing iterative exercise and that ranking of plans is the inevitable consequence of metrics analysis
- Accepting that optimization is a multivariable problem and that claims of achievements of optimization to a given target molecule are legitimate



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 15

Lautens ring construction strategy

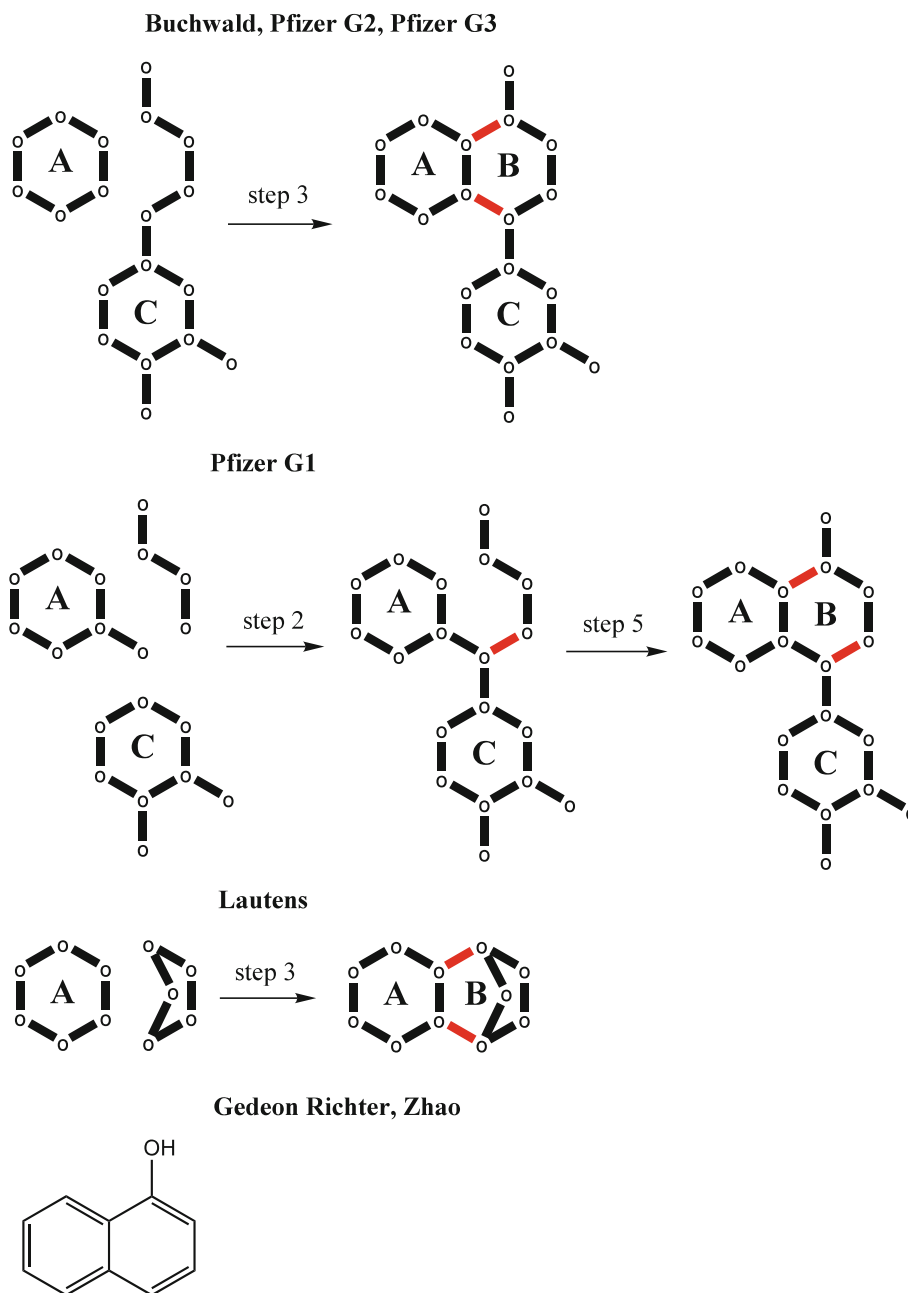
when the magnitudes of all variables in the set of plans considered are appropriately maximized and minimized and appear in the same plan

7. Realizing that when a new target molecule is desired to be made with no prior published guidance available, it is necessary to go through poor performing plans before hitting on the “right” one
8. Realizing that quantitative assessment of synthesis plans has an important role in deciding which may be good candidate plans to pursue
9. Changing the well-worn paradigm of designing synthesis plans around a fixed type of reaction just because of its novelty or because it honors the discoverer of that reaction to one that uses material and synthetic efficiency as the uppermost constraints in the choice of the set of reactions ultimately selected for a plan
10. Reaching a consensus on defining the set of first- and second-generation feedstock compounds that form a common basis of starting materials for all synthetic target molecules so that synthesis plans to any given target may be traced back to these and then ranked in an unbiased way
11. Ceasing to divorce the goals of green chemistry from those of elegant synthesis design and cost-efficient process chemistry
12. Creating a searchable structure database based on target bond mapping so that ring construction

strategies may be encoded and as an aid to ensure novelty in planning a synthesis to a given target using retrosynthetic analysis

Some caveats to bear in mind when carrying out green metrics analyses include:

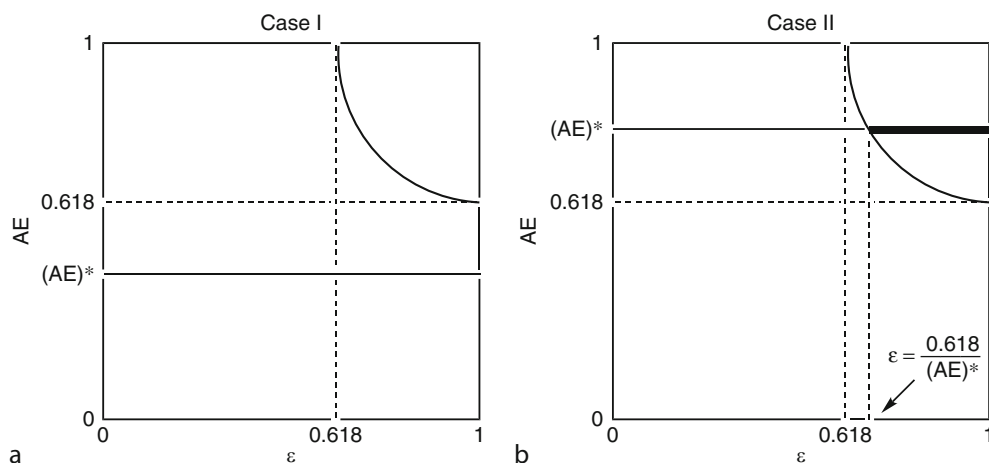
1. Generalizations about relative efficiencies of linear versus convergent routes, short plans versus longer ones, or stereoselective versus racemic with resolution must be taken with caution as there exist plenty of examples in the literature when counterintuitive results occur because gains made in one set of parameters are lost in others.
2. The synthesis plans of specialized catalysts or solvents used, such as chiral catalysts, ligands, and ionic liquids, must also be worked out using separate synthesis tree diagrams with appropriate mole-scaling factors as part of the overall assessment of material efficiency to a given target molecule.
3. When using the radial pentagon analysis for single reactions, one must be aware of determining correct mole amounts for reagents that are reported in procedures as solutions given in terms of weight percent and properly assigning the role of each material used in the right place in the template spreadsheet.
4. No claims of greenness can be made if only one plan exists for a given target molecule.



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 16
 Graphs showing ring construction strategies

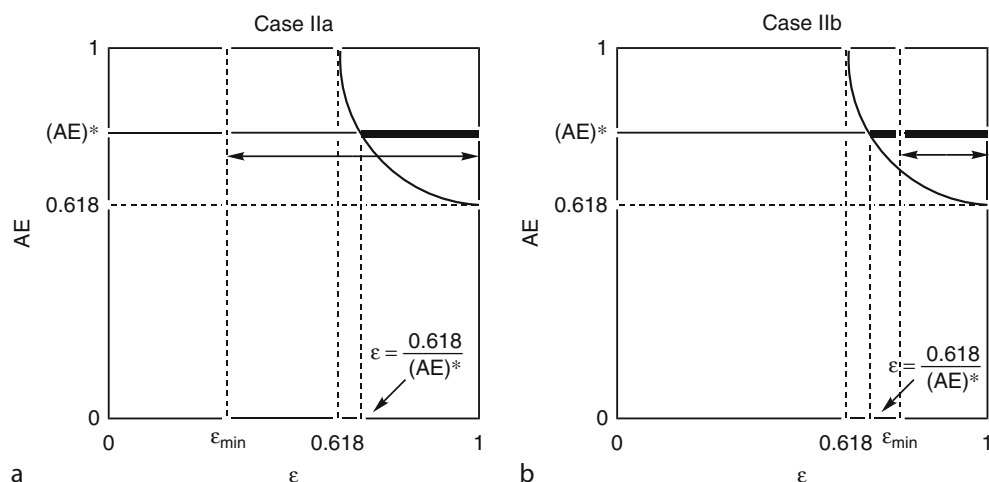
5. Claims of greenness cannot be made based on one criterion, such as the use of a “green” solvent for example.
6. Once green metrics analyses are done on a set of literature procedures to a given target molecule and

it is found in the ranking process that optimized parameters are scattered over a number of plans, it is imperative that the next disclosed plan should demonstrate improvements over the best reported prior plan.



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 17

Plots of AE versus reaction yield according to $AE = \phi/\epsilon = 0.618/\epsilon$ showing the two possible cases: **(a)** when $(AE)^* \leq 0.618$ and $0 < \epsilon_{\min} \leq 1$; and **(b)** when $0.618 < (AE)^* \leq 1$ and $0 < \epsilon_{\min} \leq 1$



Green Chemistry Metrics: Material Efficiency and Strategic Synthesis Design. Figure 18

Plots of AE versus reaction yield according to $AE = \phi/\epsilon = 0.618/\epsilon$ showing the two possible cases: **(a)** $0.618 < (AE)^* \leq 1$ and $0 < \epsilon_{\min} < \frac{0.618}{(AE)^*}$; and **(b)** when $0.618 < (AE)^* \leq 1$ and $\frac{0.618}{(AE)^*} < \epsilon_{\min} \leq 1$

Bibliography

Primary Literature

1. Andraos J (2005) Unification of reaction metrics for green chemistry: applications to reaction analysis. *Org Process Res Dev* 9:149–163
2. Andraos J (2005) Unification of reaction metrics for green chemistry II: evaluation of named organic reactions and application to reaction discovery. *Org Process Res Dev* 9:404–431
3. Andraos J (2005) On using tree analysis to quantify the material, input energy, and cost throughput efficiencies of simple and complex synthesis plans and networks: towards a blueprint for quantitative total synthesis and green chemistry. *Org Process Res Dev* 10:212–240
4. Andraos J, Izhakova J (2006) Perspectives on the application of green chemistry principles to total synthesis design. *Chimica Oggi/The Int J Ind Chem* 24(6, Suppl.):31–36
5. Andraos J, Sayed M (2007) On the use of “green” metrics in the undergraduate organic chemistry lecture and laboratory to

- assess the mass efficiency of organic reactions. *J Chem Educ* 84:1004–1010
6. Andraos J (2007) Gauging material efficiency. *Can Chem News* 59(4):14–17
 7. Trost BM (1991) The atom economy – a search for synthetic efficiency. *Science* 254:1471–1477
 8. Trost BM (2002) On inventing reactions for atom economy. *Acc Chem Res* 35:695–705
 9. Trost BM (1995) Atom economy. A challenge for organic synthesis - homogeneous catalysis leads the way. *Angew Chem Int Ed* 34:259–281
 10. Sheldon RA (1994) Consider the environmental quotient. *Chem Tech* 24(3):38–47
 11. Sheldon RA (2000) Atom utilisation, E factors and the catalytic solution. *CR Acad Sci Paris Sér IIc Chim* 3:541–551
 12. Sheldon RA (2001) Atom efficiency and catalysis in organic synthesis. *Pure Appl Chem* 72:1233–1246
 13. Curzons AD, Constable DJC, Mortimer DN, Cunningham VL (2001) So you think your process is green, how do you know? – using principles of sustainability to determine what is green - a corporate perspective. *Green Chem* 3:1–6
 14. Constable DJC, Curzons AD, Freitas dos Santos LM, Geen GR, Hannah RE, Hayler JD, Kitteringham J, McGuire MA, Richardson JE, Smith P, Webb RL, Yu M (2001) Green chemistry measures for process research and development. *Green Chem* 3:7–9
 15. Steinbach A, Winkenbach R (2000) Choose processes for their productivity. *Chem Eng April*: 94–104
 16. Constable DJC, Curzons AD, Cunningham VL (2002) Metrics to “green” chemistry – which are the best? *Green Chem* 4:521–527
 17. Eissen M, Metzger JO (2002) Environmental performance metrics for daily use in synthetic chemistry. *Chem Eur J* 8:3580–3585
 18. Eissen M, Hungerbühler K, Dirks S, Metzger J (2003) Mass efficiency as metric for the effectiveness of catalysts. *Green Chem* 5:G25–G27
 19. Metzger JO, Eissen M (2004) Concepts on the contribution of chemistry to a sustainable development - renewable raw materials. *CR Acad Sci Paris Sér IIc Chim* 7:569–581
 20. Eissen M, Mazur R, Quebbemann HG, Pennemann KH (2004) Atom economy and yield of synthesis sequences. *Helv Chim Acta* 87:524–535
 21. van Aken K, Strekowski L, Patiny L (2006) EcoScale, a semi-quantitative tool to select an organic preparation based on economical and ecological parameters. *Beilstein J Org Chem* 2. doi:10.1186/1860-5397-2-3
 22. Andraos J (2009) Global green chemistry metrics analysis algorithm and spreadsheets: evaluation of the material efficiency performances of synthesis plans for oseltamivir phosphate (Tamiflu) as a test case. *Org Process Res Dev* 13:161–185
 23. Welch WM, Kraska AR, Sarges R, Koe BK (1984) Nontricyclic antidepressant agents derived from cis- and trans-1-amino-4-aryltetraline. *J Med Chem* 27:1508–1515
 24. Quallich GJ, Williams MT, Friedmann RC (1999) Friedel-Crafts synthesis of 4-(3, 4-dichlorophenyl)-3, 4-dihydro-1(2 H)-naphthalenone, a key intermediate in the preparation of the antidepressant sertraline. *J Org Chem* 55:4971–4973
 25. Lautens M, Rovis T (1999) Selective functionalization of 1, 2-dihydronaphthalenols leads to a concise, stereoselective synthesis of sertraline. *Tetrahedron* 55:8967–8976
 26. Yun J, Buchwald SL (2000) Efficient kinetic resolution in the asymmetric hydrosilylation of imines of 3-substituted indanones and 4-substituted tetralones. *J Org Chem* 65:767–774
 27. Vukics K, Fodor T, Fischer J, Fellegvári I, Lévai S (2002) Improved industrial synthesis of antidepressant sertraline. *Org Process Res Dev* 6:82–85
 28. Taber GP, Pfisterer DM, Colberg JC (2004) A new and simplified process for preparing N-[4-(3, 4-dichlorophenyl)-3, 4-dihydro-1(2 H)-naphthalenylidene]methanamine and a telescoped process for the synthesis of (1 S-cis)-4-(3, 4-dichlorophenyl)-1, 2, 3, 4-tetrahydro-N-methyl-1-naphthalenamine mandelate: key intermediates in the synthesis of sertraline hydrochloride. *Org Process Res Dev* 8:385–388
 29. Wang G, Zheng C, Zhao G (2006) Asymmetric reduction of substituted indanones and tetralones catalyzed by chiral dendrimer and its application to the synthesis of (+)-sertraline. *Tetrahedron Asymm* 17:2074–2081
 30. Hendrickson JB (1971) A systematic characterization of structures and reactions for use in organic synthesis. *J Am Chem Soc* 93:6847–6854
- ## Books and Reviews
- Abdel-Magid FA (2004) Chemical process research: the art of practical organic synthesis. American Chemical Society, Washington
- Anastas PT, Warner JC (1998) Green chemistry: theory and practice. Oxford University Press, Oxford
- Anderson NG (2000) Practical process research and development. Academic, San Diego
- Baran PS, Maimone TJ, Richter JM (2007) Total synthesis of marine natural products without using protecting groups. *Nature* 446:404–408
- Bertz SH, Sommer TJ (1993) Applications of graph theory to synthesis planning: complexity, reflexivity, and vulnerability. In: Hudlicky T (ed) Organic synthesis: theory and applications, vol 2, JAI Press. Greenwich, CT, pp 67–92
- Burns NZ, Baran PS, Hoffmann RW (2009) Redox economy in organic synthesis. *Angew Chem Int Ed* 48:2854–2867
- Calvo-Flores FG (2009) Sustainable chemistry metrics. *ChemSusChem* 2:905–919
- Carey JS, Laffan D, Thomson C, Williams MT (2006) Analysis of the reactions used for the preparation of drug candidate molecules. *Org Biomol Chem* 4:2337–2347
- Carlson R (1992) Design and optimization in organic synthesis. Elsevier, Amsterdam
- Cornforth JW (1993) The trouble with synthesis. *Aust J Chem* 46:157–170
- Eissen M (2001) Bewertung der Umweltverträglichkeit organisch-chemischer Synthesen. PhD thesis, Universität Oldenburg

- Fuchs PL (2001) Increase in intricacy – a tool for evaluating organic synthesis. *Tetrahedron* 57:6855–6875
- Hendrickson JB (1977) Systematic synthesis design. 6. Yield analysis and convergency. *J Am Chem Soc* 99:5439–5450
- Hoffmann RW (2006) Protecting-group-free synthesis. *Synlett* 3531–3541
- Lapkin A, Constable DJC (2008) Green chemistry metrics: measuring and monitoring sustainable processes. Wiley, Chichester
- Lee S, Robinson G (1995) Process development: fine chemicals from grams to kilograms. Oxford University Press, Oxford
- Newhouse T, Baran PS, Hoffmann RW (2009) The economies of synthesis. *Chem Soc Rev* 38:3010–3021
- Nicolaou KC, Vourloumis D, Winssinger N, Baran PS (2000) The art and science of total synthesis at the dawn of the twenty-first century. *Angew Chem Int Ed* 39:44–122
- Nicolaou KC (2003) Perspectives in total synthesis: a personal account. *Tetrahedron* 59:6683–6738
- Nicolaou KC, Snyder SA (2004) The essence of total synthesis. *Proc Nat Acad Sci USA* 101:11929–11936
- Nicolaou KC, Edmonds DJ, Bulger PG (2006) Cascade reactions in total synthesis. *Angew Chem Int Ed* 45:7134–7186
- Orru RVA, de Greef M (2003) Recent advances in solution-phase multicomponent methodology for the synthesis of heterocyclic compounds. *Synthesis* 1471–1499
- Posner GH (1986) Multicomponent one-pot annulations forming three to six bonds. *Chem Rev* 86:831–844
- Qiu F (2008) Strategic efficiency – the new thrust for synthetic organic chemists. *Can J Chem* 86:903–906
- Seebach D (1990) Organic synthesis – where now? *Angew Chem Int Ed* 29:1320–1367
- Serratosa F (1990) Organic chemistry in action: the design of organic synthesis. Elsevier, Amsterdam
- Sheldon RA (1997) The E factor: fifteen years on. *Green Chem* 9:1273–1283
- Sheldon RA (2008) E factors, green chemistry and catalysis: an odyssey. *Chem Commun* 3352–3365
- Smit WA, Bochkov AF, Caple R (1998) Organic synthesis: the science behind the art. Royal Society of Chemistry, Cambridge
- Snieckus V (1999) Optimization in organic synthesis. *Med Res Rev* 19:342–347
- Tietze LF (1996) Domino reactions in organic synthesis. *Chem Rev* 96:115–136
- Tietze LF, Modi A (2000) Multicomponent domino reactions for the synthesis of biologically active natural products and drugs. *Med Chem Rev* 20:304–322
- Ugi I, Dömling A, Hörl W (1994) Multicomponent reactions in organic chemistry. *Endeavour New Ser* 18(3):115–122
- Ugi I, Dömling A, Werner B (2000) Since 1995 the new chemistry of multicomponent reactions and their libraries, including their heterocyclic chemistry. *J Heterocycl Chem* 37:647–658
- Ugi I (2001) Recent progress in the chemistry of multicomponent reactions. *Pure Appl Chem* 73:187–191
- Weber L, Illgen K, Almstetter M (1999) Discovery of new multicomponent reactions with combinatorial methods. *Synlett* 366–374
- Weber L (2002) Multi-component reactions and evolutionary chemistry. *Drug Discov Today* 7:143–147
- Weber L (2002) The application of multi-component reactions in drug discovery. *Curr Med Chem* 9:1241–1253
- Wender P, Miller BL (1993) Toward the ideal synthesis: connectivity analysis and multibond-forming processes. In: Hudlicky T (ed) *Organic synthesis: theory and applications*, vol 2, JAI Press. Greenwich, Connecticut, pp 27–65
- Zhang TY (2006) Process chemistry: the science, business, logic, and logistics. *Chem Rev* 106:2583–2595
- Zhu J, Bienyamé H (2005) Multicomponent reactions. Wiley, Weinheim

Green Chemistry with Microwave Energy

RAJENDER S. VARMA

Sustainable Technology Division, National Risk Management Research Laboratory, U.S. Environmental Protection Agency, Cincinnati, OH, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Reactions Under Solvent-Free Conditions (With or Without Any Support)
 Reactions in Aqueous Medium
 Future Directions
 Disclaimer
 Bibliography

Glossary

Green chemistry Green chemistry is the broad discipline that encompasses the design of chemical processes and products that eliminate or reduce the generation and use of hazardous substances. It applies across the life cycle, including the design, manufacture, and use of a chemical product.

Microwaves Microwaves (0.3–300 GHz) lie in the electromagnetic radiation spectrum between radiowave (Rf) and infrared (IR) frequencies with relatively large wavelengths and are a form of energy and not heat. This nonionizing radiation, incapable of breaking chemical bonds, is a form of energy that

manifests itself as heat through interaction with the polar medium.

Sustainability Literally meaning to “maintain,” “support,” or “endure” the concept of sustainability calls for policies and strategies that meet society’s present needs without compromising the ability of future generations to meet their own needs.

Definition of the Subject

Green chemistry utilizes a set of 12 principles that reduces or eliminates the use or generation of hazardous substances in the design, manufacture, and applications of chemical products [1]. This newer chemical approach protects the environment by inventing safer and eco-friendly chemical processes that prevent pollution “at source” rather than cleaning up “end-of-the-pipe” by-products and pollutants generated by traditional synthesis. The diverse nature of our chemical universe promotes a need for various greener strategic pathways in our quest to attain sustainability. The synthetic chemical community has been under increased pressure to produce, in an environmentally benign fashion, the myriad of chemical entities required by society in relatively short spans of time. This is especially true for the pharmaceutical and fine chemical industries. Among others, one of the best options is to accelerate these synthetic processes by using microwave (MW)-assisted chemistry techniques in conjunction with safer reaction media. The efficient use of the MW heating approach for the synthesis of a wide variety of organics and nanomaterials in aqueous and solvent-free media is discussed in this entry, including the sustainable application of recyclable and reusable nano-catalysts.

Introduction

The emerging area of green chemistry emphasizes minimum hazard as the performance criteria while designing new chemical processes. Rather than remediation, which involves cleaning up of waste after it has been produced, the main objective is to avoid waste generation in the first place. The desired approach requires new environmentally benign syntheses, catalytic methods, and chemical products that are “benign by design” [1]. One of the thrust areas for achieving this target is to explore alternative expeditious reaction

conditions and eco-friendly reaction media to accomplish the desired chemical transformations with minimized by-products or waste as well as eliminating the use of conventional organic solvents. Consequently, several newer strategies have appeared, such as solvent-free (dry media) reactions with [2, 3] and without microwave irradiation [4].

The nonclassical heating technique using microwaves, termed as the “Bunsen burner of the 21st century,” is rapidly becoming popular for expeditious chemical syntheses and transformations [5]. This entry summarizes noteworthy greener methods that use MWs that have resulted in the development of sustainable synthetic protocols for drugs and fine chemicals. A brief account of the author’s own experiences in MW-assisted organic transformations involving benign alternatives, such as solid-supported reagents [2, 4], and greener reaction media, such as aqueous, ionic liquid, and solvent-free, for the synthesis of various heterocycles [3], oxidation-reduction reactions, coupling reactions, and some name reactions is included [5–9].

Microwaves are a nonionizing form of radiation energy that are not strong enough to break chemical bonds but transfer energy selectively to various substances. Some materials (such as hydrocarbons, glass, and ceramics) are nearly transparent to microwaves and therefore behave as good insulators in a MW oven since they are heated only to a very limited extent. Metals reflect MW; molecules with dipole moment (many types of organic compounds) and salts absorb MW energy directly. Microwaves couple directly with molecules in the reaction mixture with rapid rise in temperature; dipole rotation and ionic conduction are the two most important fundamental mechanisms for transfer of energy from microwaves to the molecules being heated. Polar molecules try to align themselves with the rapidly changing electric field of the MW and the coupling ability is determined by the polarity of the molecules. Therefore, there are some significant differences in terms of thermal gradient between the conventional chemical reactions in the liquid phase and the same reactions conducted under MW irradiation.

When a liquid reaction mixture is subjected to conventional heating, the walls of the vessel are directly heated but the reaction mixture receives thermal energy by conduction/convection. The temperature of the

reaction mixture cannot be higher than the temperature of the vessel walls. In contrast, MW energy permeates through glass vessels directly and is available for absorption by molecules in the reaction mixture. Consequently, it is possible for these molecules to be at higher energy levels and the contents of the flask to be at a higher bulk temperature in a few minutes. Another important feature of microwaves is that they penetrate several centimeters into a liquid; in contrast, radiant heat (e.g., from infrared rays) raises the temperature of only the surface layer. A reaction mixture under MW irradiation is therefore at a higher temperature in the middle than at the surface. Accordingly, the temperature profile of the reaction mixture can be quite different depending on whether there is conventional heating or MW-induced energy transfer.

Reactions Under Solvent-Free Conditions (With or Without Any Support)

Although the initial surface-mediated chemical transformation dates back to 1924 [10], it was not until almost half a century later that the technique received well-deserved attention as attested by several books [11, 12, 13], reviews, and account articles [14, 15]. Heterogeneous reactions facilitated by supported reagents on inorganic oxide surfaces have received special attention in recent years. The use of MW irradiation techniques for the acceleration of organic reactions had a profound impact on these heterogeneous reactions since the appearance of initial reports on the application of microwaves for chemical synthesis in polar solvents [16, 17]. The approach has now matured into a useful technique for a variety of applications in organic synthesis and functional group transformations, as is testified by a large number of books [18, 19] and review articles on this theme [2–5, 20–23].

The reactions appear to occur at relatively low bulk temperature although higher localized temperatures may be reached during MW irradiation. Unfortunately, accurate recording of temperature has not been made in the majority of such studies. The situation has now improved as a result of the availability of commercial MW systems. The recyclability of some of these solid mineral supports renders them into environmentally friendlier “green” protocols.

Since the initial report [24], a large number of MW-promoted solvent-free protocols have been illustrated for a wide variety of useful chemical transformations, such as cleavage (protection/deprotection), condensation, rearrangement reactions, oxidation, reduction, and the synthesis of several heterocyclic compounds on mineral supports [2–5, 20–23]. A range of industrially significant chemical precursors, such as enones, imines, enamines, nitroalkenes, and heterocyclic compounds, has been obtained in a relatively environmentally friendlier manner [2–5, 20–23]. A vast majority of these solvent-free reactions has been performed using an unmodified household MW oven or commercial MW equipment usually operating at 2,450 MHz in open glass containers with neat reactants. The general procedure involves simple mixing of neat reactants with the catalyst, their adsorption on mineral or “doped” supports, and subjecting the reaction mixture to MW irradiation.

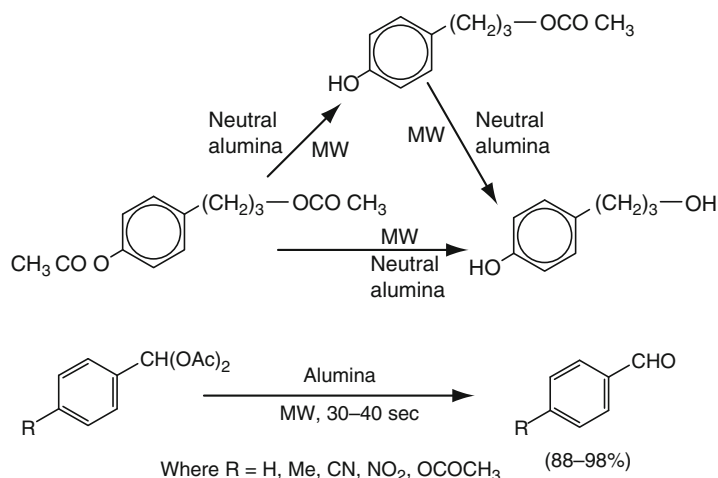
It is clear that this solventless approach addresses the problems associated with waste disposal of solvents that are used severalfold in chemical reactions, thus minimizing or avoiding the excess use of chemicals and solvents.

Protection–Deprotection Reactions

Although inherently wasteful, protection–deprotection reaction sequences constitute an integral part of organic syntheses, such as the preparation of monomers, fine chemicals, and precursors for pharmaceuticals. These reactions often involve the use of acidic, basic, or hazardous reagents, and toxic metal salts [25]. The solventless MW-accelerated cleavage of functional groups provides an attractive alternative to conventional deprotection reactions.

Deacylation Reactions The utility of recyclable alumina as a viable support surface for deacylation reaction was reported by Varma and his colleagues wherein the orthogonal deprotection of alcohols [26] and regeneration of arylaldehydes from the corresponding diacetates [27] is possible under solvent-free conditions on neutral alumina surface using MW irradiation (Scheme 1).

Debenzylation of Carboxylic Esters An efficient solvent-free debenzylation process for the cleavage of



Green Chemistry with Microwave Energy. Scheme 1

Deacylation of protected alcohols and phenols and regeneration of aldehydes from aldehyde diacetates on neutral alumina

carboxylic esters on alumina surface was developed by Varma and coworkers [28]. By changing the surface characteristics of the solid support from neutral to acidic, the cleavage of 9-fluorenylmethoxycarbonyl (Fmoc) group and related protected amines was achieved. The cleavage of *N*-protected moieties, however, required the use of basic alumina and a MW irradiation time of 12–13 min at ~130–140°C.

Desilylation Reactions *Tertiary*-butyldimethylsilyl (TBDMS) ether derivatives of alcohols can be rapidly cleaved to regenerate the corresponding hydroxy compounds on alumina using MW irradiation [29]. This approach circumvents the use of corrosive fluoride ions that is normally employed for cleaving such silyl protecting groups.

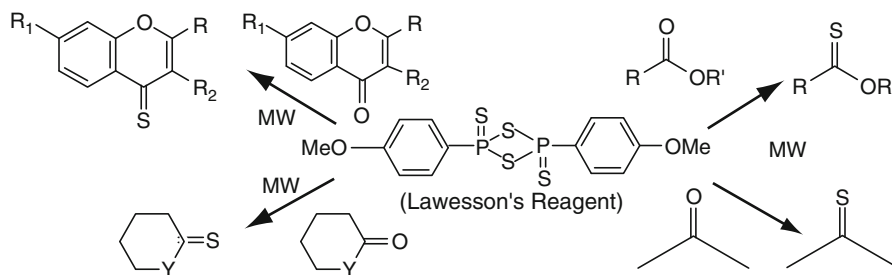
Thionation Reactions: Thioketones, Thiolactones, Thioamides, Thionoesters, and Thioflavonoids The conversion of carbonyl compounds to the corresponding thio analogues is especially useful under solventless conditions and circumvents the use of conventional phosphorous pentasulfide under basic conditions, hydrogen sulfide in the presence of acid, or Lawesson's reagent. Using the MW approach, no acidic or basic medium is used and the carbonyl compounds are simply admixed with neat Lawesson's

reagent (0.5 equiv.). This benign approach is general and is applicable to the high yield conversion of ketones, flavones, isoflavones, lactones, amides, and esters to the corresponding thio analogs (Scheme 2). The protocol uses comparatively a much smaller amount of Lawesson's reagent and avoids the use of large excesses of dry hydrocarbon solvents, such as benzene, xylene, triethylamine, or pyridine, that are conventionally used [30].

Dethioacetalization Thio acetals and ketals derived from aldehydes and ketones were rapidly deprotected by clayfen within seconds under solvent-free conditions (Scheme 3), thus avoiding the use of excess solvents and toxic oxidants commonly employed in the dethioacetalization process [31].

Oxidation Reactions

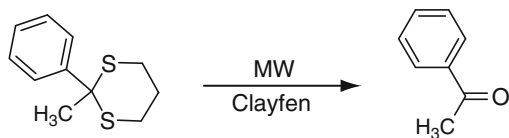
Oxidation of Alcohols The oxidation of alcohols to carbonyl compounds is an important transformation in organic synthesis. The use of supported reagents has gained popularity because of the improved selectivity, reactivity, and associated ease of manipulation. Alcohols can be rapidly and selectively oxidized to the corresponding carbonyl compounds by silica-supported active manganese dioxide (Scheme 4) under solvent-free conditions using microwaves [32].



Where Y = O, NH and R, R', R₁ and R₂ are aryl or alkyl groups

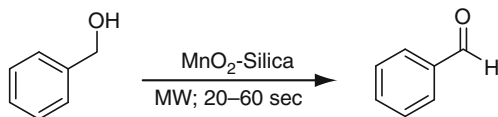
Green Chemistry with Microwave Energy. Scheme 2

Solvent-free synthesis of thioketones, thiolactones, thioamides, and thionoesters by Lawesson's reagent. From Varma and Kumar (Ref. [30])



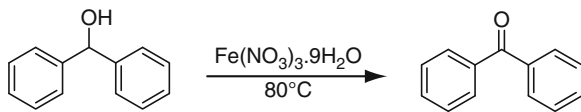
Green Chemistry with Microwave Energy. Scheme 3

Clayfen-catalyzed dethioacetalization



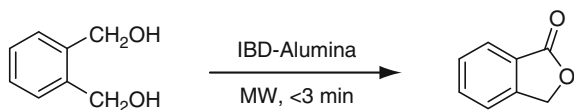
Green Chemistry with Microwave Energy. Scheme 4

MnO₂-silica-catalyzed oxidation of alcohols



Green Chemistry with Microwave Energy. Scheme 5

Solventless oxidation of alcohols to carbonyl compounds



Green Chemistry with Microwave Energy. Scheme 6

IBD-alumina catalyzed oxyhyperiodination of alcohols

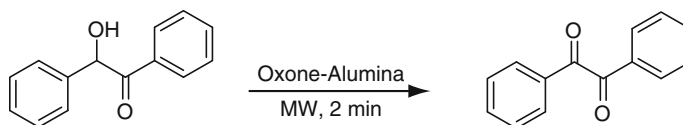
This oxidation protocol can also be catalyzed by ferric nitrate under solvent-free conditions [33], without any solid support (Scheme 5).

The nonmetallic oxidant, iodobenzene diacetate (IBD) on alumina, accomplished the oxidation of alcohols under solvent-free MW irradiation conditions and entails only the mixing of the neat alcohols with 1.1 equivalents of IBD doped on neutral alumina (Scheme 6); this rapid procedure avoids the over-oxidation of alcohols to carboxylic acids [34].

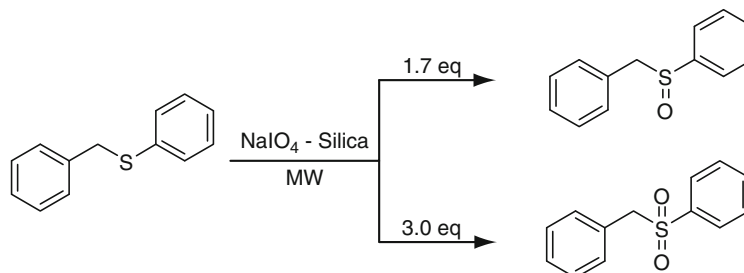
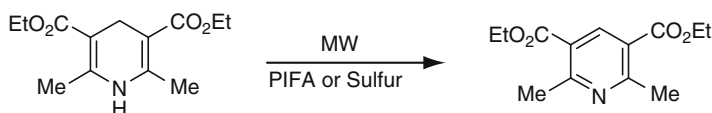
The solid-state oxidation of α -hydroxyketones to generate vicinal ketones (benzils) is possible using Oxone[®] supported on wet alumina (Scheme 7). The protocol circumvents the major limitations of Oxone[®], namely, the low solubility in organic solvents and the inherent danger of combustibility upon heating with solvents [35].

Oxidation of Sulfides Selective oxidation of sulfides to sulfoxides and sulfones is possible [36] using wet silica-supported sodium periodate under MW irradiation (Scheme 8). A unique feature of this protocol is its applicability to long-chain aliphatic sulfides which are usually insoluble in polar solvents and are difficult to oxidize.

Oxidation of Dihydropyridines Solventless oxidation of 1,4-dihydropyridines to pyridines was effected by phenyliodine(III) bis(trifluoroacetate) (PIFA) at room temperature or elemental sulfur under MW irradiation conditions (Scheme 9). Dealkylation at the 4-position in the cases of ethyl, isopropyl, and

**Green Chemistry with Microwave Energy. Scheme 7**

Oxone-alumina catalyzed oxidation of benzoin

**Green Chemistry with Microwave Energy. Scheme 8**NaIO₄-silica catalyzed selective oxidation of sulfides**Green Chemistry with Microwave Energy. Scheme 9**

PIFA or sulfur-catalyzed oxidation of 1,4-dihydropyridines

benzyl-substituted dihydropyridine derivatives with PIFA was circumvented by an alternative general procedure using elemental sulfur which provides pyridines in good yields [37].

Reduction Reactions

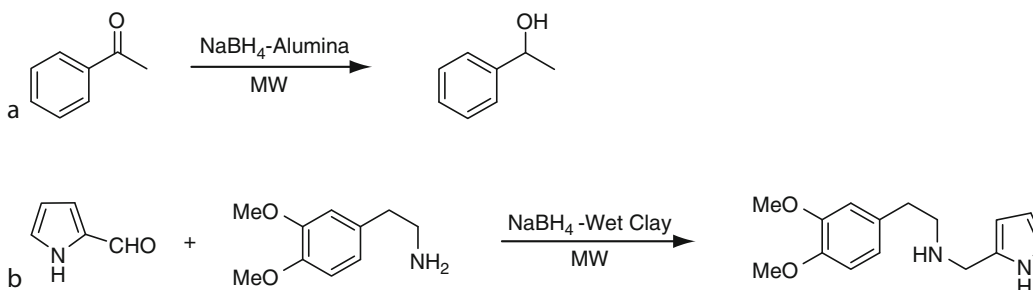
Reduction of Carbonyl Compounds A manipulative simple and rapid method for the reduction of carbonyl compounds was developed [38] by Varma et al. under solvent-free conditions using NaBH₄-Alumina and MW irradiation (Scheme 10a). When sodium borohydride on wet clay was used as support, the reductive-amination of carbonyl compounds was efficiently accomplished under MW irradiation (Scheme 10b) [39]. The reactions involving Schiff's bases generated from cyclohexanone and aniline and aliphatic aldehydes and amines, however, require a relatively longer time for completion.

Reduction of Nitro Compounds Varma and coworkers developed a solvent-free MW reduction protocol that leads to a facile preparation of aromatic amines from the corresponding nitro compounds [40] with hydrazine hydrate supported on solid materials such as alumina, silica gel, and clays (Scheme 11).

Similarly, hydroxylamine supported on clay can convert arylaldehydes into nitriles (Scheme 12) [41].

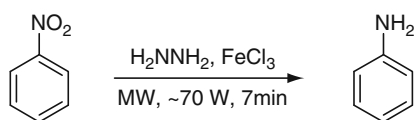
MW-Assisted Synthesis of Ionic Liquids and Their Application as Catalysts

Ionic liquids (ILs) have been the focus of attention due to their potential in a variety of commercial applications, such as electrochemistry, heavy metal ion extraction, phase transfer catalysis, and polymerization, and as substitutes for conventional volatile organic solvents (Rogers and Seddon 2002) [42, 43]. Additional environmentally friendly attributes of these ILs include



Green Chemistry with Microwave Energy. Scheme 10

(a) NaBH_4 -catalyzed reduction of carbonyl compounds and (b) Reductive-amination of carbonyl compounds



Green Chemistry with Microwave Energy. Scheme 11

Reduction of nitro compounds to amines

negligible vapor pressure, potential for recycling, compatibility with various organic compounds and organometallic catalysts, and ease of separation of products from reactions.

Solvent-Free Synthesis of Ionic Liquids Ionic liquids, being polar and ionic in character, couple with MW irradiation very efficiently and are, therefore, ideal MW-absorbing candidates for enhancing chemical reactions. The first efficient preparation of 1,3-dialkylimidazolium halides via MW irradiation was developed by Varma et al. (Scheme 13) [44, 45]. The reaction time was reduced to minutes and the method avoids the excessive use of organic solvents as the reaction medium. These syntheses can also be ideally carried out using ultrasound under solvent-free conditions [46].

Metal-bearing classes of ILs, $[\text{Rmim}][\text{InCl}_4]$ and $[\text{bmim}][\text{GaCl}_4]$, were prepared using a solvent-free MW procedure (Scheme 14) and utilized as catalysts [47, 48]. This approach is much faster, efficient, and eco-friendly as it does not use any organic solvent.

Synthetic Application of ILs as Catalysts Ionic liquids have emerged as a new class of green solvents for chemical processes and transformations (Rogers and

Seddon 2002). Their polarity renders them good solvents for various organic reactions and catalyzes, including the dissolutions of renewable materials such as cellulose [49].

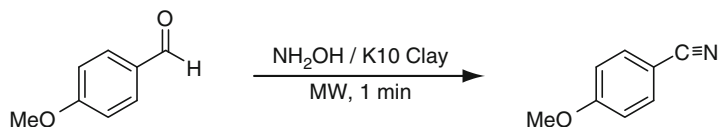
1-Butyl-3-methylimidazolium tetrachlorogallate, $[\text{bmim}][\text{GaCl}_4]$, has been used as an active catalyst for the efficient acetalization of aldehydes (Scheme 15) [48].

Imidazolium-based tetrachloroaluminates ($[\text{Rmim}][\text{AlCl}_4]$) [50] and tetrachloroindate ($[\text{Rmim}][\text{InCl}_4]$) [47] have been used as recyclable catalysts for the efficient protection of alcohols to form tetrahydropyranyl (THP) derivatives (Scheme 16).

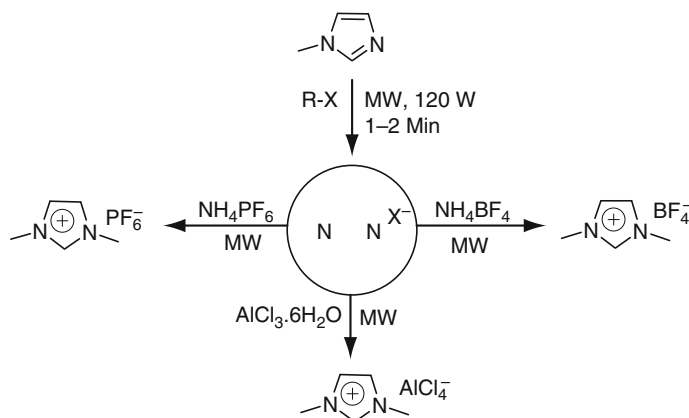
The coupling reaction of carbon dioxide (CO_2) with epoxides generates the five-membered cyclic carbonates which are precursors for polymeric materials such as polyurethanes and polycarbonates, serve as aprotic polar solvents, and are utilized as intermediates in the production of pharmaceutical and fine chemicals. The catalytic amounts of tetrahaloindate (III)-based ILs are found to exhibit the highest catalytic activities for the synthesis of cyclic carbonates (Scheme 17) [51].

Reactions in Aqueous Medium

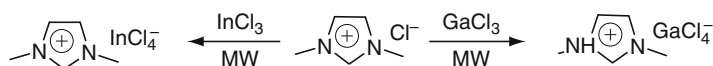
Solvent-free reactions undoubtedly minimize the environmental impacts resulting from the use of solvents in chemical production but there are limitations in view of the poorly understood heat- and mass-transfer issues under these conditions [2]. Fluorous solvents, ionic liquids [52], aqueous systems [7], and supercritical carbon dioxide have emerged as alternatives in this movement. Water, which is naturally abundant and can

**Green Chemistry with Microwave Energy. Scheme 12**

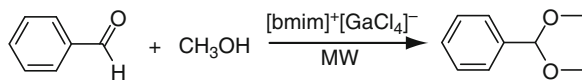
Hydroxylamine-clay-catalyzed synthesis of aromatic nitriles

**Green Chemistry with Microwave Energy. Scheme 13**

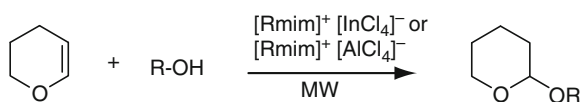
MW-assisted synthesis of ionic liquids

**Green Chemistry with Microwave Energy. Scheme 14**

MW-assisted synthesis of In and Ga containing ionic liquids

**Green Chemistry with Microwave Energy. Scheme 15**

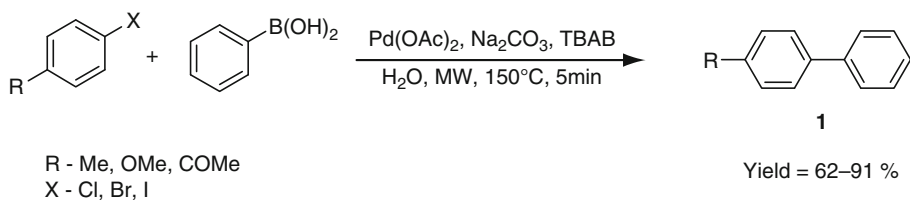
Ionic liquid-catalyzed protection of carbonyl compounds

**Green Chemistry with Microwave Energy. Scheme 16**

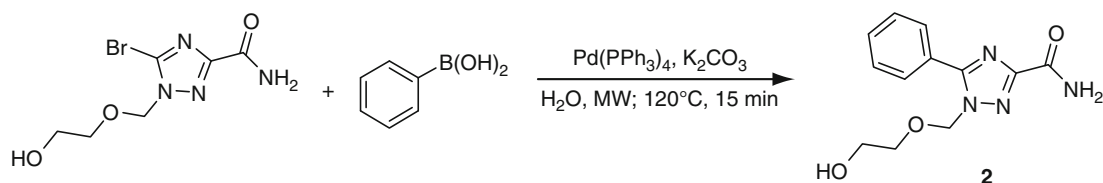
Ionic liquid-catalyzed THP-protection of alcohols

be contained because of its relatively higher vapor pressure, appears to be a sustainable alternative because of its nontoxic, noncorrosive, and nonflammable nature [6–8].

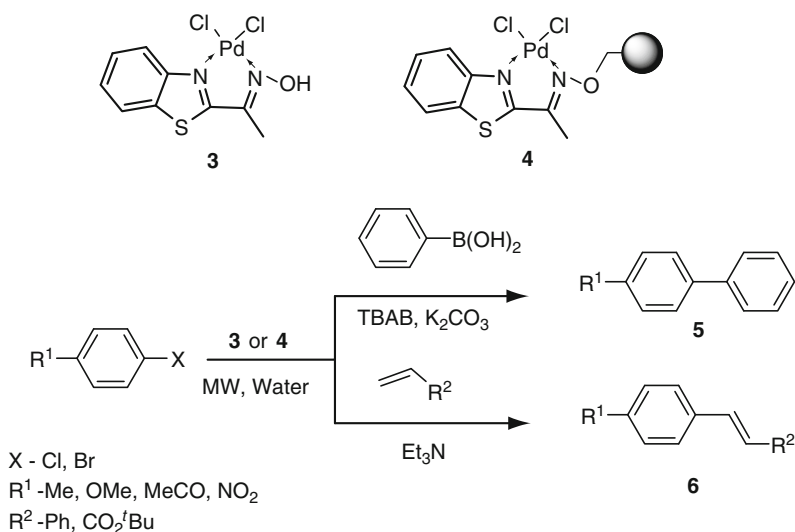
In addition to the isolation of products from the aqueous medium, the insolubility of most of the organic substrates in water is a challenge. However, some of these issues can be overcome by using the MW heating technique wherein water is rapidly heated to high temperatures enabling it to act like a pseudo-organic solvent. Two key mechanisms, namely, dipolar polarization and ionic conduction of water molecules (Fig. 1), come into play upon irradiation of a reaction

**Green Chemistry with Microwave Energy. Scheme 18**

Aqueous MW-assisted Suzuki reaction

**Green Chemistry with Microwave Energy. Scheme 19**

Aqueous MW-assisted Suzuki reaction of triazole

**Green Chemistry with Microwave Energy. Scheme 20**

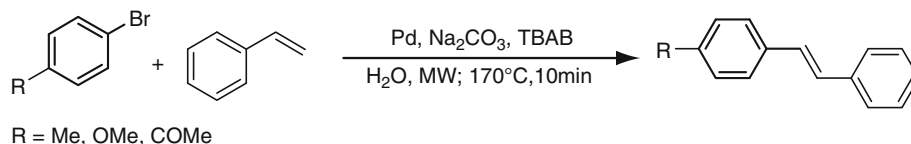
Aqueous MW Suzuki and Heck reactions using benzothiazole-based Pd(II) complexes

Arvela and Leadbeater performed the Heck coupling reaction in water using MW heating (Scheme 21), with Pd catalyst concentrations as low as 500 ppb [60].

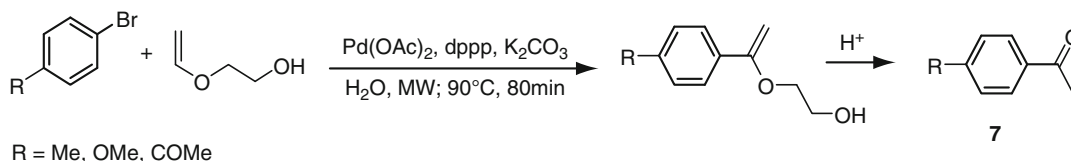
Larhed et al. have recently reported highly regioselective and fast Pd(0)-catalyzed internal R-arylation

of ethylene glycol vinyl ether with aryl halides in aqueous medium (Scheme 22) in presence of 1,3-bis(diphenylphosphino)propane (dppp) [61].

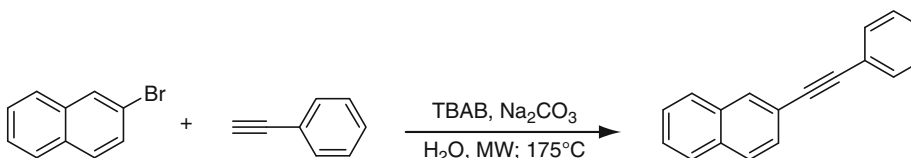
Aryl bromides and iodides were efficiently converted to corresponding acetophenones [7] in high yields in water using ethylene glycol vinyl ether as the olefin and

**Green Chemistry with Microwave Energy. Scheme 21**

Aqueous MW-assisted Heck reaction

**Green Chemistry with Microwave Energy. Scheme 22**

Aqueous MW-expedited internal Heck reaction

**Green Chemistry with Microwave Energy. Scheme 23**

Aqueous MW-assisted Sonogashira reaction

potassium carbonate as the base. This Pd(0)-catalyzed method is advantageous as no heavy metal additives or ionic liquids are necessary. The reaction proceeded cleanly without any noticeable byproduct formation and avoided the need for inert atmosphere. MW irradiation has proven to be beneficial in activation of aryl chlorides toward the internal Heck arylation.

An aqueous Sonogashira coupling reaction was reported by Eycken et al. under MW irradiation (Scheme 23). This reaction precludes the need for copper(I) or any transition-metal phosphane complex, thus overcoming the problem of toxicity and air sensitivity of transition-metal complexes, as well as the use of expensive phosphane ligands [62].

Hiyama Cross-Coupling Reaction

The relatively benign and stable organosilanes have emerged as useful entities for cross-coupling reactions.

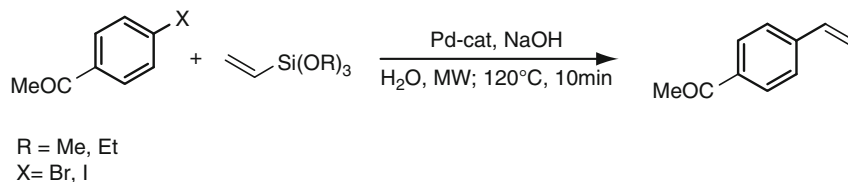
This is exemplified by the Hiyama cross-coupling reactions between vinylalkoxysilanes and aryl halides which were promoted by aqueous sodium hydroxide under fluoride-free conditions and carried out using MW irradiation in aqueous medium (Scheme 24) [63].

Stille Reaction

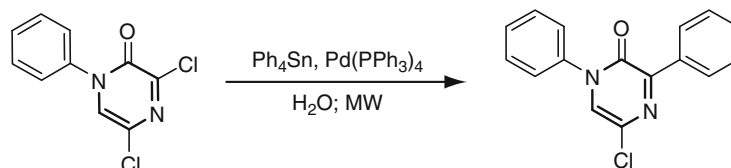
Stille reaction between organo-tin and aryl halide was achieved using aqueous MW chemistry by van der Eycken et al. (Scheme 25) [64].

Transformation of Amines to Ketones

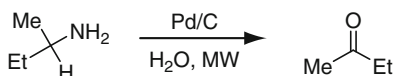
Although the selective transformation of amines to ketones is a common biological process, conversion of amines to ketones by chemical means are rather limited. A retro-reductive, MW-assisted, Pd-catalyzed amination reaction for direct conversion of amines to

**Green Chemistry with Microwave Energy. Scheme 24**

Aqueous MW-assisted Hiyama reaction

**Green Chemistry with Microwave Energy. Scheme 25**

Aqueous MW-assisted Stille reaction

**Green Chemistry with Microwave Energy. Scheme 26**

MW-assisted aqueous retro-reductive amination reaction

ketones in water was accomplished by Olah et al. (Scheme 26). This expeditious reaction proceeds smoothly without any heavy metal-based oxidant or volatile organic solvents [65].

Decarboxylation of Cinnamic Acids

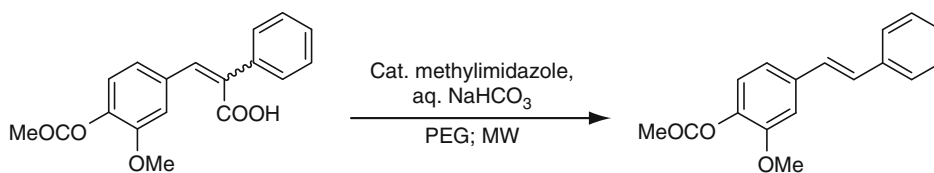
A metal-free protocol for decarboxylation of substituted α -phenylcinnamic acid derivatives in aqueous media was developed by Sinha et al. [66], wherein a catalytic amount of methylimidazole in aqueous NaHCO_3 in polyethylene glycol (PEG) under MW irradiation conditions furnished the corresponding para/ortho hydroxylated (E)-stilbenes (Scheme 27). This is clearly a clean alternative to the conventional multistep methods that often involve the use of toxic quinoline and a copper salt combination. The critical role of water in facilitating the decarboxylation highlights the synthetic utility of water-mediated organic transformations.

Reactions in Near-Critical Water

Dallinger and Kappe examined several MW-assisted organic reactions in near-critical water (NCW) in the temperature range of 270–300°C. The hydrolysis of esters or amides, the hydration of alkynes, Diels–Alder cycloadditions, pinacol rearrangements, and the Fischer indole synthesis were successfully performed in MW-generated NCW without the addition of an acid or base catalyst [67]. This study demonstrated that it is technically feasible to perform MW synthesis in water on scales from 15 to 400 ml at temperatures of up to 300°C and 80 bars of pressure in a multimode MW reactor.

Synthesis of Heterocycles

Heterocyclic compounds are a special class among pharmaceutically significant natural products and synthetic compounds [68, 69]. The remarkable ability of heterocyclic nuclei to serve both as biomimetics and reactive pharmacophores has provided a unique value as traditional key elements for numerous drugs. Conventional organic synthesis is too slow to satisfy the demand for generation of small molecules and this void has been filled by combinatorial and automated chemistry to meet the increasing requirement of new



Green Chemistry with Microwave Energy. Scheme 27

Decarboxylation of cinnamic acid in aqueous medium using microwaves

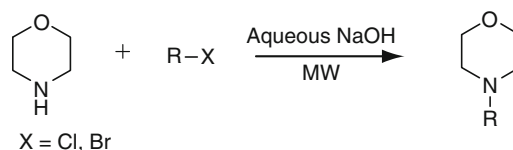
compounds for drug discovery, expediently [70]. The efficiency of MW flash-heating has resulted in dramatic reductions in reaction times from days and hours to minutes and seconds [5, 6, 9, 71–73].

Nitrogen Heterocycles Nitrogen heterocycles are abundant in nature and are of great significance to life because their structural subunits exist in many natural products such as vitamins, hormones, antibiotics, and alkaloids, as well as herbicides, dyes, and pharmaceuticals [68].

An expedient *N*-alkylation of nitrogen heterocycles was reported in aqueous media under MW irradiation conditions (Scheme 28) [74].

The double *N*-alkylation of primary amines and hydrazine derivatives (Scheme 29) in carbonated water using microwaves provided access to the synthesis of nitrogen-containing heterocycles, namely, substituted azetidines, pyrrolidines, piperidines, azepanes, *N*-substituted 2,3-dihydro-1H-isoindoles, 4,5-dihydropyrazoles, pyrazolidines, and 1,2-dihydro-phthalazines. The readily available alkyl dihalides (or ditosylates) thus provide a facile entry to important classes of building blocks in natural products and pharmaceuticals [75–77].

This MW-accelerated general approach shortened the reaction time significantly and utilized readily available amines and hydrazines with alkyl dihalides or ditosylates to assemble two C–N bonds in a simple S_N2 -like sequential heterocyclization experimental protocol which has never been fully realized under conventional reaction conditions. The strategy avoids multistep reactions, functional group protection/deprotection sequences, and eliminates the use of expensive phase transfer and transition metal catalysts. The experimental observations are consistent with the mechanistic postulation



Green Chemistry with Microwave Energy. Scheme 28

N-Alkylation in basic water using MW irradiation

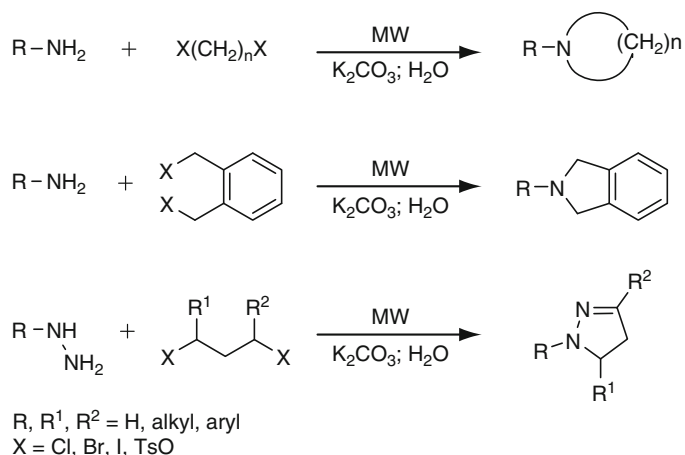
wherein the polar transition state of the reaction is favored by MW irradiation with respect to the dielectric polarization nature of MW energy transfer. In large-scale experiments, the phase separation of the desired product in either solid or liquid form from the aqueous media can facilitate product purification by simple filtration or decantation instead of tedious column chromatography, distillation, or extraction processes, which reduces the usage of volatile organic solvents [77].

A direct Grignard type of addition of alkynes to in situ generated imines, from aldehyde and amines, was catalyzed by CuBr and provides a rapid and solvent-free approach access to propargylamines in excellent yields (Scheme 30) [78].

Direct synthesis of these cyclic ureas using a MW-assisted protocol that proceeds rapidly in the presence of ZnO were reported (Scheme 31) [79].

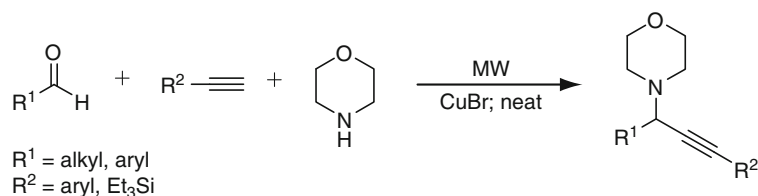
The imidazo[1,2-*a*] annulated nitrogen heterocycles bearing pyridine, pyrazine, and pyrimidine moieties constitute a class of biologically active compounds that can be assembled via atom-economic, one-pot condensation of aldehydes, amines, and isocyanides (three-component Ugi reaction) using a MW approach (Scheme 32) in the presence of recyclable montmorillonite K-10 clay under solvent-free conditions [80].

Dihydropyrimidinones are an important class of biologically active organic compounds that were synthesized under solvent-free conditions [81] or by an



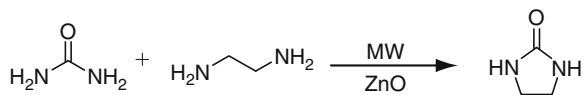
Green Chemistry with Microwave Energy. Scheme 29

Synthesis of *N*-heterocycles in aqueous media using MW irradiation



Green Chemistry with Microwave Energy. Scheme 30

CuBr-catalyzed solvent-free synthesis of propargylamines



Green Chemistry with Microwave Energy. Scheme 31

ZnO-catalyzed MW synthesis of imidazolidine-2-one

environmentally benign Biginelli protocol using polystyrene sulfonic acid (PSSA) as a catalyst (Scheme 33). A very simple eco-friendly isolation procedure entails the filtration of the precipitated products [82].

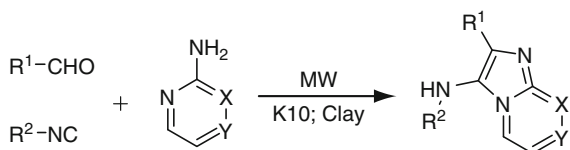
Oxygen-Heterocycles Heterocycles bearing oxygen atom are important classes of building blocks in organic synthesis and have attracted the attention of medicinal chemists over the years [69].

The solvent-free synthesis of 2-aminosubstituted isoflav-3-enes was developed, which can be carried

out in one pot using microwaves via the in situ generation of enamines and their subsequent reactions with salicylaldehydes (Scheme 34) [83]. This environmentally friendly procedure does not require azeotropic removal of water using large excesses of aromatic hydrocarbon solvents for the generation of enamines or the activation of the catalyst.

The expeditious solvent-free syntheses of 2-aryloxybenzo[*b*]furans were developed from readily accessible α -tosyloxyketones and mineral oxides in processes that are accelerated by exposure to microwaves (Scheme 35) [84].

Dioxane rings, common structural motifs in numerous bioactive molecules, can be assembled via tandem bis-aldol reaction of ketones with paraformaldehyde in aqueous media catalyzed by PSSA under MW irradiation conditions to produce 1,3-dioxanes (Scheme 36) [85].



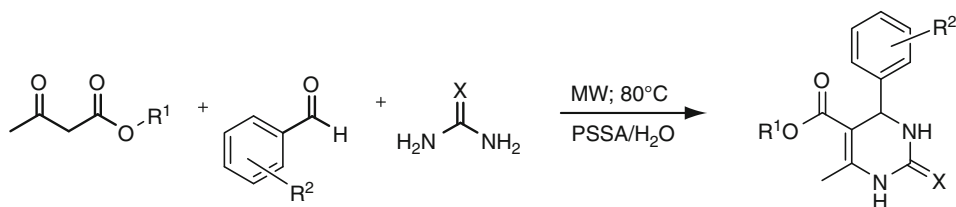
X, Y = C or N

Green Chemistry with Microwave Energy. Scheme 32

Clay-catalyzed solvent-free synthesis of annulated *N*-heterocycles

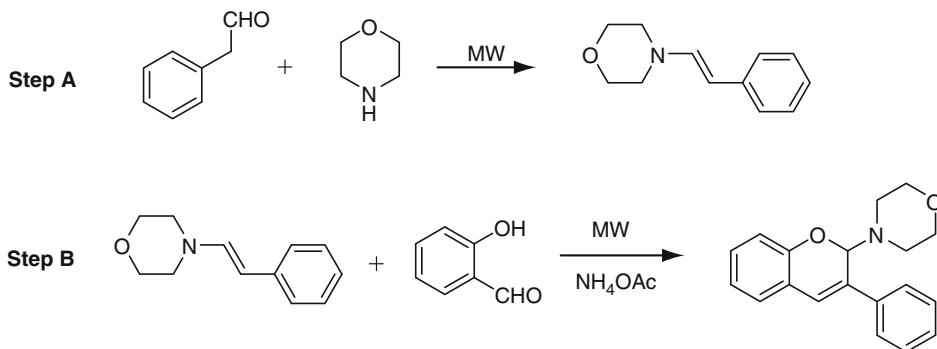
Heterocyclic Hydrazones The first example of a reaction between two solids, under solvent-free and catalyst-free environment, was accomplished by Varma et al. The reaction of neat 5- or 8-oxobenzopyran-2(1H)-ones with a variety of aromatic and heteroaromatic hydrazines provides rapid access to several synthetically useful heterocyclic hydrazones (Scheme 37) [86].

An aqueous protocol for the synthesis of heterocyclic hydrazones using PSSA as a catalyst was also devel-



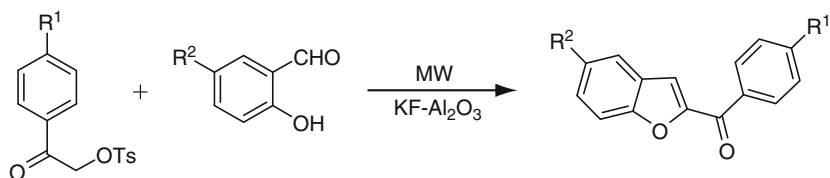
Green Chemistry with Microwave Energy. Scheme 33

Biginelli reaction using microwaves in aqueous medium



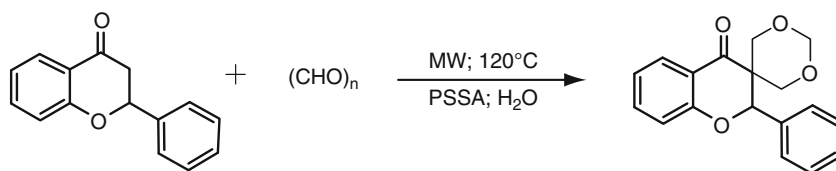
Green Chemistry with Microwave Energy. Scheme 34

One-pot solventless synthesis of isoflav-3-enes



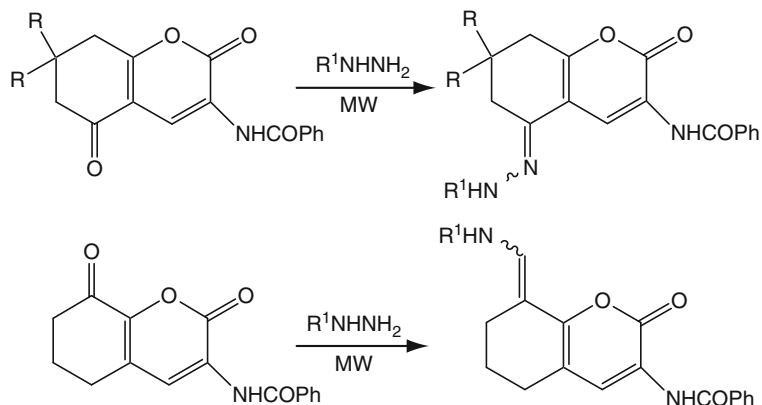
Green Chemistry with Microwave Energy. Scheme 35

MW assisted synthesis of 2-arylbenzo[*b*]furans



Green Chemistry with Microwave Energy. Scheme 36

PSSA-catalyzed synthesis of 1,3-dioxanes in aqueous media



Green Chemistry with Microwave Energy. Scheme 37

Synthesis of heterocyclic hydrazones under solvent-free conditions (Ref. [85])

oped by this group recently (Scheme 38). The protocol entails the simple filtration as the product isolation step in a reaction that proceeds in the absence of any organic solvent under MW irradiation [87].

Sulfur Heterocycles Synthesis of thiazoles often involves utilization of lachrymatory starting materials, phenacyl halides, and hazardous reagents under drastic conditions. Varma and coworkers accomplished the synthesis of 1,3-thiazoles (which are not easily accessible under classical heating conditions) in excellent yields [83] from thioamides and α -tosyloxyketones catalyzed by montmorillonite K-10 clay, (Scheme 39).

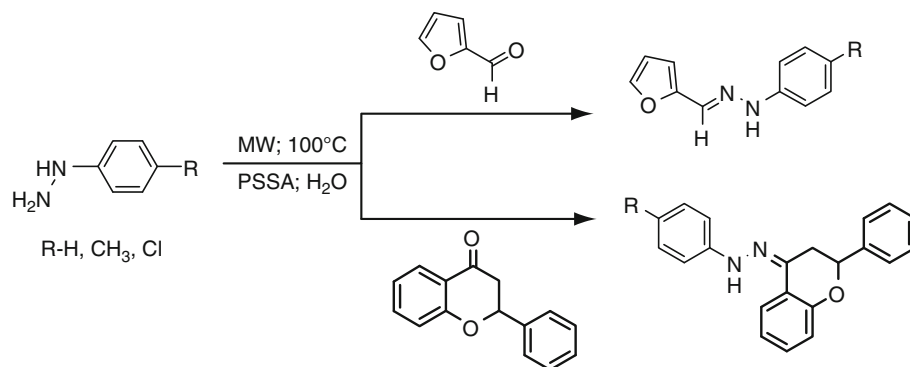
The strategy was extended to a concise synthesis of bridgehead 3-aryl-5,6-dihydroimidazo[2,1-*b*][1,3]thiazoles [8], which are difficult to obtain, requiring heating over an extended period of time, and uses α -haloketones or α -tosyloxyketones under strongly acidic conditions. The solventless mixing of α -tosyloxyketones with thioamides in the presence of

montmorillonite K-10 clay and exposing the reactants to MW irradiation for 3 min affords the substituted bridgehead thiazoles (Scheme 40) [83].

A novel one-pot solventless synthesis of 1,3,4-oxadiazoles and 1,3,4-thiadiazoles via condensation of acid hydrazide and triethyl orthoalkanates under MW irradiation was recently developed (Scheme 41) [88].

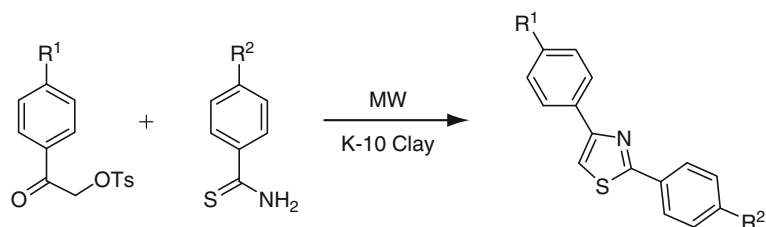
Synthesis of Nanomaterials

The use of MW irradiation as an efficient, environmentally friendly, and economically viable heating method for the production of nanomaterials has increased in recent years. Synthesis of gold (Au) nanoparticles with various morphologies using sugar as reducing agent under MW heating was reported by Nadagouda and Varma [89]. The authors reported the synthesis of nanoparticles (Fig. 2), prism, hexagons, and rods using a wide variety of sugars (*D*-glucose, sucrose, mannose, etc.) as reducing agents. The highly dispersed



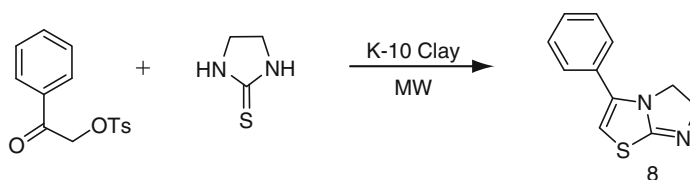
Green Chemistry with Microwave Energy. Scheme 38

Synthesis of hydrazone derivatives of furaldehyde and flavanone in water



Green Chemistry with Microwave Energy. Scheme 39

MW-assisted solventless synthesis of 1,3-thiazoles



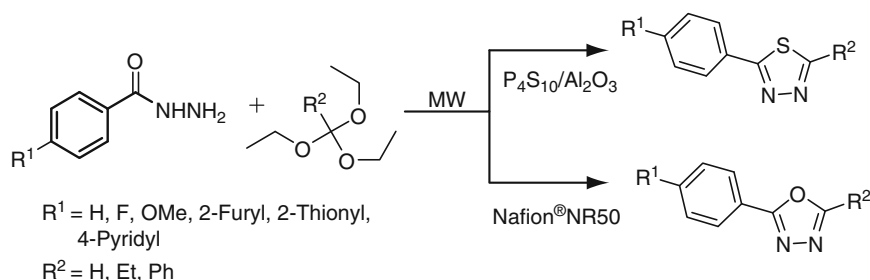
Green Chemistry with Microwave Energy. Scheme 40

MW-assisted solventless synthesis of bridgehead-thiazoles

nanoparticles are produced in the size range of 10–20 nm in less than a minute. When low concentration of sugar is used, hexagonal, triangular, and rod-shaped particles with submicron sizes are obtained.

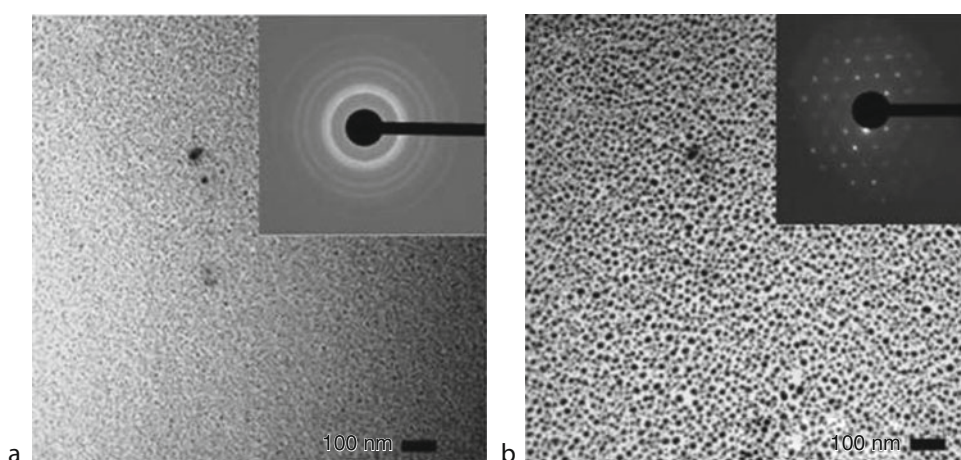
Baruwati and Varma [90] reported the MW-assisted synthesis of Au nanoparticles along with

other noble metal nanoparticles using red wine and red grape pomace extract as a single source of reducing and capping agent; highly dispersed nanoparticles are produced in the size range 5–20 nm (Fig. 3) and only 45–60 s reaction time is required. This general method produces silver (Ag), Pd, platinum (Pt), and iron (Fe)



Green Chemistry with Microwave Energy. Scheme 41

One-pot solventless synthesis of 1,3,4-oxadiazoles and 1,3,4-thiadiazoles

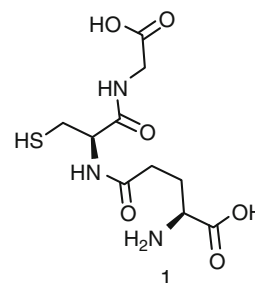


Green Chemistry with Microwave Energy. Figure 2

Au nanoparticles synthesized with (a) glucose and (b) sucrose under MW irradiation conditions (Ref. [88])

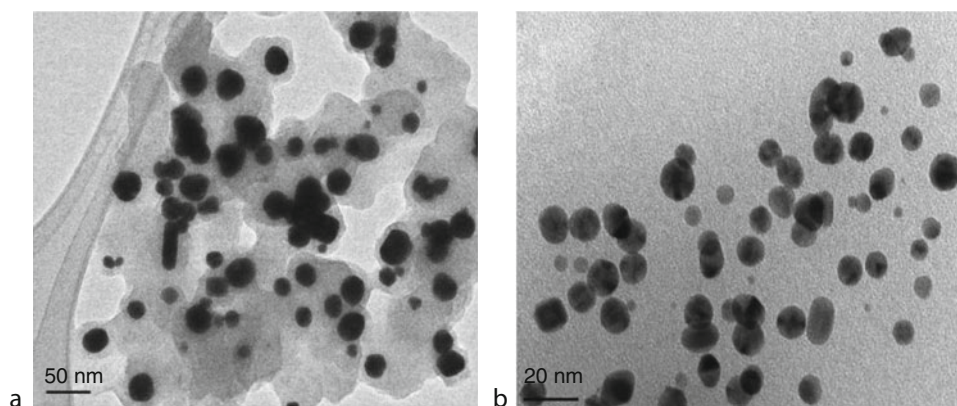
in the particle range 5–20 nm. In contrast, white wine produces highly agglomerated nanoparticles because of lack of polyphenolics which are present in red grape pomace as well as in red wine.

Highly dispersed Ag nanoparticles in the size range 5–10 nm were prepared by Varma et al. [91] using aqueous MW approach within a minute utilizing a completely benign and ubiquitous tripeptide, glutathione (Structure 1). Normally, crystalline Ag nanoparticles (Fig. 4) are obtained and the effect of MW power on the morphology of ensuing Ag nanoparticles has also been investigated for this green and sustainable procedure which is adaptable for the synthesis of Pd, Pt, and Au nanoparticles



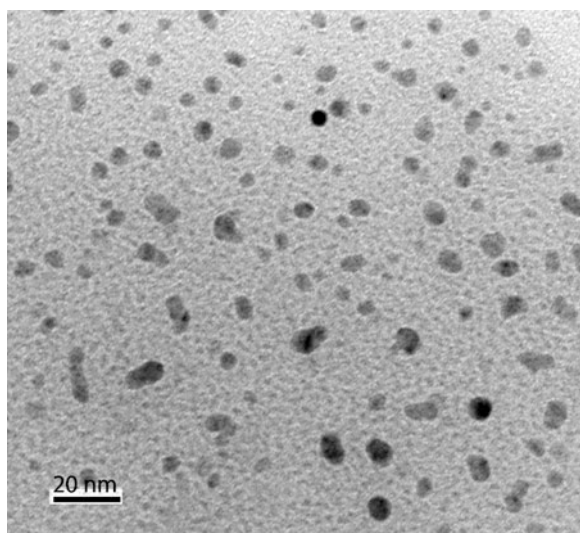
Chemical structure of tripeptide, Glutathione

Greener Synthesis of Nanomaterials The development of solution-based controlled synthesis of nanomaterials via a bottom-up approach often uses



Green Chemistry with Microwave Energy. Figure 3

Au nanoparticles synthesized using (a) red wine and (b) red grape pomace at MW power level 50 W, time 1 min (Ref. [89])



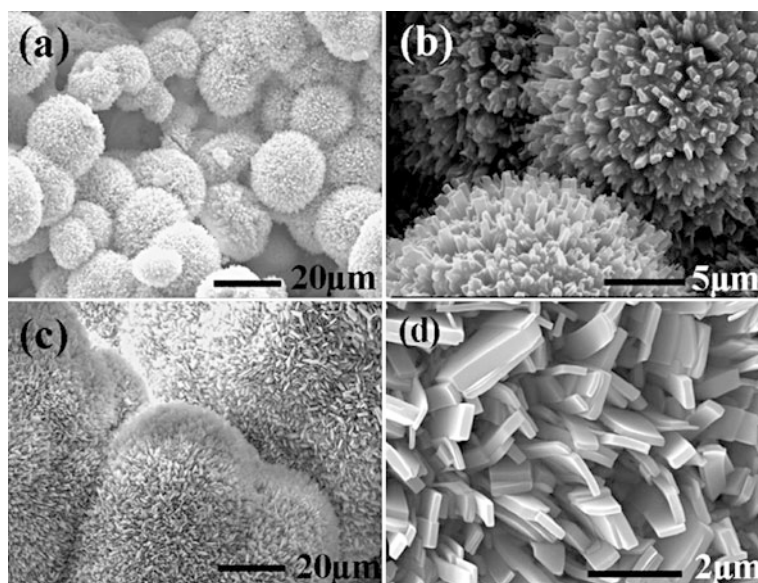
Green Chemistry with Microwave Energy. Figure 4

TEM micrograph of Ag nanoparticles at MW power level of 50 W for 1 min (Ref. [90])

toxic reducing and capping agents and some dispersants. Greener alternatives, especially using a biomimetic approach, are now possible wherein benign entities such as vitamin B₁ [92], vitamin B₂ [93], vitamin C [94], tea polyphenols [95], simple sugars [89], and PEG [96] can generate nanoparticles. These nanoparticles can be cross-linked under the influence of microwaves to form nanocomposites with cellulose [97] or polyvinyl alcohol (PVA) [98].

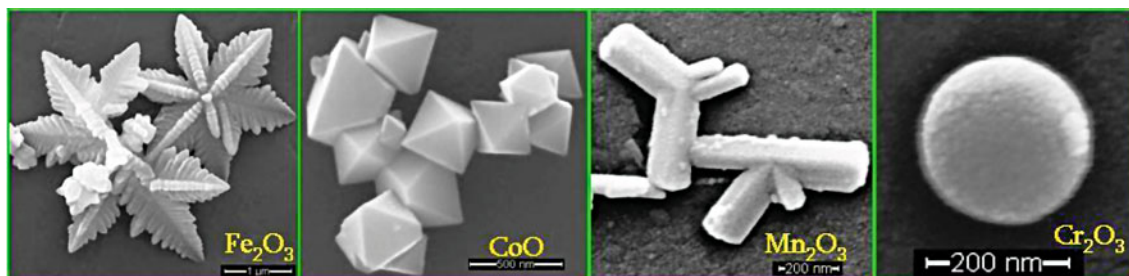
Vitamin B₁ was used for one-step synthesis of Pd nanobelts, nanoplates, and nanotrees without using any special capping agents (Fig. 5) [92]. Depending upon the Pd concentration, Pd nanoparticles crystallized in various shapes and sizes. At lower Pd concentration, plate-like shape was obtained. The Pd plates were grown on a single Pd nanorod backbone mimicking the leaf-like structures. However with increase in Pd concentration, formation of tree-like structures was observed. Upon further increase in concentration, Pd nanoplates become thicker by vertically aligning themselves to form ball-like shape and this general protocol can be extended to prepare other noble nanomaterials such as Au and Pt. The Pd nanoparticles showed excellent catalytic activity for several C–C bond-forming reactions such as Suzuki, Heck, and Sonogashira reactions under MW irradiation conditions [92].

The control of the size and morphology of nanostructures to tailor the physical and chemical properties is an important aspect in nanoscience. Recently, Varma et al. designed a convenient method for the synthesis of metal oxides with 3D nanostructures [99, 100] which are obtainable from hydrolysis of inexpensive starting materials in water without using any reducing or capping reagent. This economical and environmentally sustainable synthetic concept could ultimately enable the fine-tuning of material responses to magnetic, electrical, optical, and mechanical stimuli. Well-defined morphologies, including octahedron, sphere, triangular rod, pine, and



Green Chemistry with Microwave Energy. Figure 5

SEM images of Pd nanoparticles generated using vitamin B₁ (Ref. [91])



Green Chemistry with Microwave Energy. Figure 6

MW-assisted synthesis of metal oxides with well-defined morphologies (Ref. [99])

hexagonal snowflake with particles in the size range of 100–500 nm were obtained (Fig. 6). Nano-ferrites were then functionalized and coated with Pd metal, which catalyzed various C–C coupling and hydrogenation reactions with high yields. In addition, the effortless recovery and increased efficiency, combined with the inherent stability of this catalyst, rendered the method sustainable [99, 100]. In view of these unique morphologies, synthesized nanomaterials will have significant applications in biomedical science and catalysis.

MW-Assisted Synthesis of Quantum Dots (QD) in Aqueous Medium During the last decade, synthesis

of high-quality semiconductor nanocrystals popularly referred to as quantum dots, QD, has been a subject of intense research because of their size-dependent properties. These colloidal semiconductor nanoparticles are much superior compared to the conventional dye molecules in terms of flexible photoexcitation, sharp photoemission, and superb resistance to photobleaching [101]. Their optical properties could be manipulated by changing the size and composition to meet specific wavelength requirements [102] and they have found application in quantum-dot lasers, optoelectronics, nonlinear optical devices, solar cells, and bio-tagging [103–108]. Synthesis of water-soluble ZnSe nanocrystals in

aqueous medium under MW conditions has been reported by Qian et al. [109]. These nanocrystals are water soluble, have high crystallinity, and their photoluminescence (PL) quantum yield ranges up to 17%. These properties are marked improvements when compared to ZnSe QDs prepared by conventional aqueous synthesis method. The method has eliminated the use of expensive, environmentally unfriendly reagents such as trioctylphosphine (TOP), tributylphosphine (TBP), and trioctylphosphine oxide (TOPO).

ZnS nanoballs were synthesized in saturated water solutions under MW irradiation conditions by Zhao et al. [110] under ambient air; the ensuing products were highly crystalline and about 300 nm in diameter. CdSe, PbSe, and $\text{Cu}_2\text{-xSe}$ nanoparticles were prepared under MW irradiation conditions by Zhu et al. [111] wherein CdSO_4 , $\text{Pb}(\text{Ac})_2$, and CuSO_4 reacted with Na_2SeSO_3 in water in the presence of complexing agents: potassium nitrilotriacetate ($\text{N}(\text{CH}_2\text{-COOK})_3\text{-NTA}$) for CdSe and PbSe or triethanol amine for $\text{Cu}_2\text{-xSe}$ in a MW refluxing system. Although the method is simple and adequate for producing CdSe nanoparticles in 4–5 nm range, PbSe, and $\text{Cu}_2\text{-xSe}$ nanoparticles were found to be bigger in size in the range 30–80 nm and often agglomerated. The method also describes how different phases of CdSe could be obtained by varying the MW heating times. A 10 min irradiation led to the formation of CdSe in the cubic (sphalerite) phase, while after 30 min, CdSe obtained is in the hexagonal cadmoselite phase.

A one-pot synthetic method for the synthesis of CdSe/ZnS core/shell QDs using MW radiation was reported by Schumacher et al. [112] and is based on the addition of a water-soluble Zn^{2+} complex, $\text{Zn}(\text{NH}_3)_4^{2+}$ to a solution containing CdSe initial nanocrystals and 3-mercaptopropionic acid (MPA). Subsequent MW heating for less than 2 h generated high-quality CdSe/ZnS-based QDs possessing good photoluminescent quantum yield (13%) and biocompatibility.

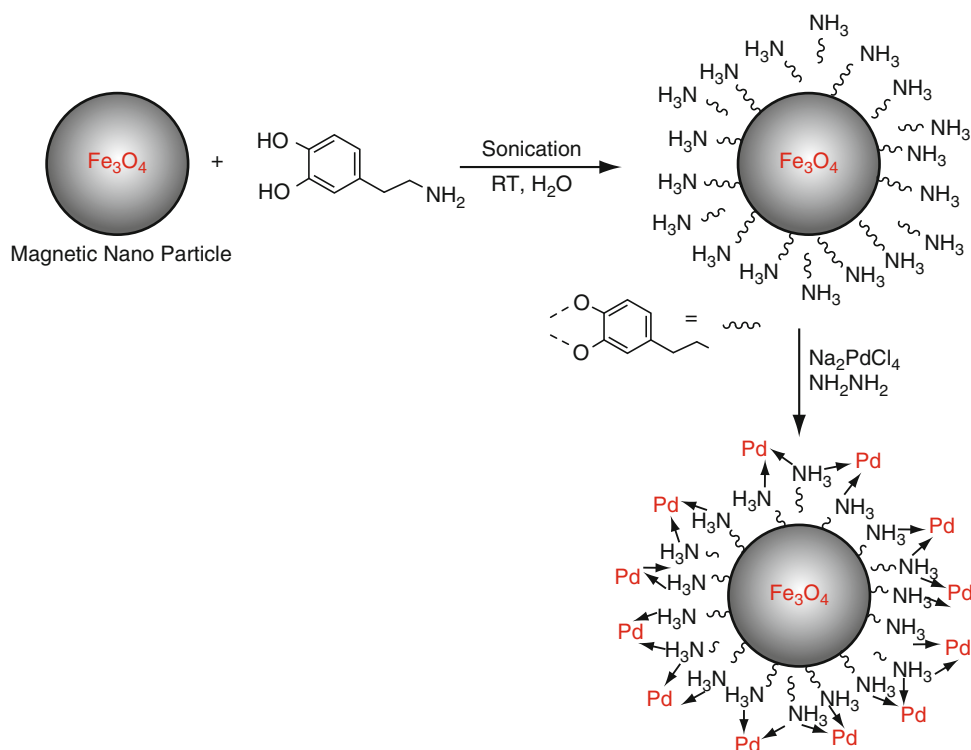
A seed-mediated approach for rapid synthesis of high-quality alloyed quantum dots (CdSe–CdS) in aqueous phase by MW irradiation was reported by Qian et al. [113]. Initially, CdSe seeds were first formed by the reaction of NaHSe and Cd^{2+} , and then alloyed quantum dots (CdSe–CdS) were rapidly generated by the release of sulfide ions from 3-mercaptopropionic acid as sulfide source with MW irradiation.

Magnetic Nanoparticles Recently, magnetic nanoparticles have emerged as viable alternatives to conventional materials, as a readily available, robust, high-surface-area heterogeneous catalyst support [114]. Post-synthetic surface modification of magnetic nanoparticles controls desirable chemical functionality and enables the generation of catalytic sites on the surfaces of resulting nano-catalyst. Their insoluble character together with paramagnetic nature enables effortless separation of these nano-catalysts from the reaction mixture using an external magnet, which eliminates the necessity of catalyst filtration. These novel nano-catalysts bridge the gap between homogeneous and heterogeneous catalysis, thus preserving the desirable attributes of both the systems.

This concept was recently explored for the development of other metal catalysts [99, 100, 115–119]. Varma and coworkers developed a convenient synthesis of nano-ferrite-supported Pd catalyst from inexpensive starting materials in water (Scheme 42) [115, 116]. This catalyst catalyzes the oxidation of alcohols and olefins with high turnover numbers and excellent selectivity. Also, being magnetically separable, this approach eliminates the requirement of catalyst filtration after completion of the reaction, which is an additional sustainable attribute of this oxidation protocol.

MW-Assisted Nano-Catalysis in Water Chemists have been under intense pressure to develop newer methods, which are expeditious and environmentally benign. One of the better alternatives is the use of nano-catalysis in conjunction with MW heating technology. The efficiency of MW heating has resulted in dramatic reductions in reaction times, reduced from days to minutes, which is very significant in process chemistry for the expedient generation of organics and nanomaterials [5, 6, 121–125].

Naturally abundant water is a good alternative because of its nonflammable, nontoxic, and noncorrosive nature, which may help reduce the dependence of chemists on hazardous solvents [6–8, 67, 74–77, 85, 126, 127]. Additionally, water can be contained because of its relatively lower vapor pressure when compared to organic solvents, thus rendering it a sustainable alternative. Interestingly, the combination of MW and aqueous medium has shown excellent benefits such as shorter reaction times, homogeneous



Green Chemistry with Microwave Energy. Scheme 42

Dopamine functionalized nanoferrite-Pd-catalyst. From Polshettiwar and Varma (Ref. [119])

in-core heating, and enhanced yields and selectivity [6–8, 67, 74–77]. In addition to these microwave “thermal effects” and “nonthermal effects” [128], there are additional benefits of using microwaves for nano-catalyzed aqueous protocols as described below:

1. Selectivity toward water

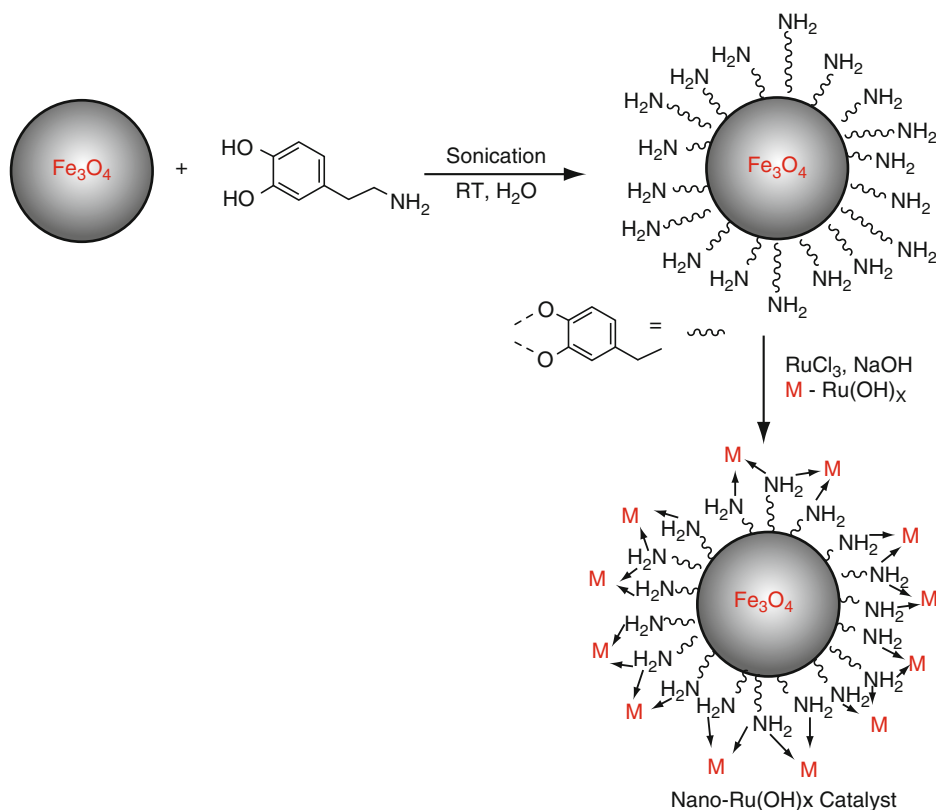
Microwave heating depends on the composition and structure of molecules (i.e., their dielectric properties) and this property can facilitate selective heating. Microwaves initiate the rapid and intense heating of polar molecules such as water, while non-polar molecules do not absorb the radiation and are not heated. Loupy and Varma [121], Strauss [122], and Larhed [129] demonstrated that this selective heating can be exploited to develop a high yield rapid MW protocol using a two phase (polar–non-polar) solvent system. The advantageous use of water in the MW-assisted processes, especially without the use of phase-transfer catalysts [121], has been well demonstrated [6–8, 67, 74–77].

2. Selectivity toward catalyst

Selective heating can be exploited in heterogeneous catalysis methods as demonstrated in MW-assisted rapid molybdenum-catalyzed allylic reactions by Larhead and his coworkers [130] and in the case of oxidation of alcohol using Magtrieve™ by Bogdal et al. [131]. These authors established that the polar catalyst absorbed extra energy and heated at a higher temperature than the overall reaction temperature, thus making the process more energy efficient.

3. Nano-catalysts serve as susceptors

Susceptors are materials that efficiently absorb MW irradiation and transfer the generated thermal energy to molecules in the vicinity that are weak MW absorbers. Although transmission of the energy occurs through conventional mechanisms, MW heating is more rapid than conventional heating. Kappe [132] and Leadbeater [133] used silicon carbide and ionic liquid, respectively, as susceptors and established that addition of these

Synthesis of Nano-Ru(OH)_x Catalyst

Green Chemistry with Microwave Energy. Scheme 43

Magnetically separable nano-Ru(OH)_x catalyst. From Polshettiwar and Varma (Ref. [119])

materials in the reaction mixture enhanced its overall capacity to absorb microwaves and significantly reduced the required MW energy. The use of these materials as susceptors in the reaction mixture, however, adds to the overall cost of the protocol. Ideally, if suitably designed nanomaterials can play a dual role of catalyst and susceptor, then the advantageous attributes can be enjoyed without the need of any additional material as a susceptor.

4. Nano-catalyst stability

MW-assisted reactions are often fast and consequently the residence time of nano-catalysts at this elevated temperature is kept to a minimum. Catalytic processes with short reaction times thus safeguard the catalyst from deactivation and decomposition, thereby increasing the overall efficiency of the catalyst and the entire method.

It appears that this approach of unifying MW technique with nano-catalysis and benign water (as a reaction medium) can offer an extraordinary synergistic effect with greater potential than these three individual components in isolation. To illustrate the concept of this green and sustainable approach, some representative protocols are presented below.

Ruthenium Hydroxide Nano-Catalyst in MW-Assisted Hydration of Nitriles in Water Amides have been generally prepared by the hydration of nitriles, with catalysis by strong acids and bases which produces several by-products including carboxylic acids. Under the influence of strong reagents and harsh conditions, however, sensitive functional groups on molecules could not be kept intact, thus decreasing the selectivity of the reaction protocol. Heterogeneous

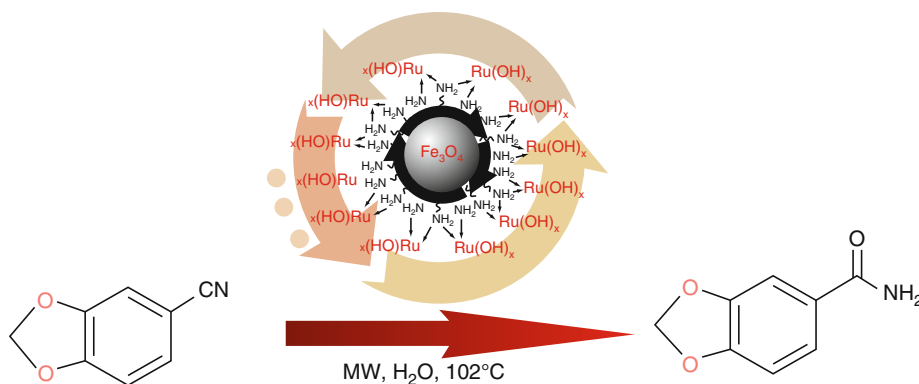
catalyst systems have been developed to overcome these drawbacks associated with homogeneous processes. The limited turnover numbers of these protocols and reusability of the catalyst are continuing challenges. A hydration method in pure water improved the reaction conditions and product yield [134], but it used expensive ruthenium (Ru) complexes as catalysts and required traditional workup using volatile organic solvents for isolation purposes. A green and sustainable pathway was developed using ruthenium hydroxide nano-catalyst under aqueous MW conditions [135] wherein nano-Ru(OH)_x was prepared in two steps. Magnetic nanoparticles were functionalized post-synthetically [100, 117] via sonication of nano-ferrites with dopamine in aqueous medium, followed by the addition of ruthenium (Ru) chloride and subsequent hydrolysis using sodium hydroxide solution (Scheme 43).

The nano-Ru(OH)_x catalyst exhibited high activity for hydration of a range of activated, inactivated, and heterocyclic nitriles in water medium and the reactions rates were not influenced by the nature of the substituents on the benzonitrile molecules. This protocol with high catalytic activity displayed excellent chemoselectivity and neither an electronic effect nor the position of the substituents influenced the reaction rate. In the hydration of the benzonitrile-containing dioxole ring, the reaction proceeded only at the cyano group to afford the corresponding amide, while keeping the ring intact (Scheme 44). Therefore, this protocol could be very useful in the total synthesis of

drug molecules, where the selective hydration of a nitrile group to an amide is a requirement, without influencing other sensitive functional groups.

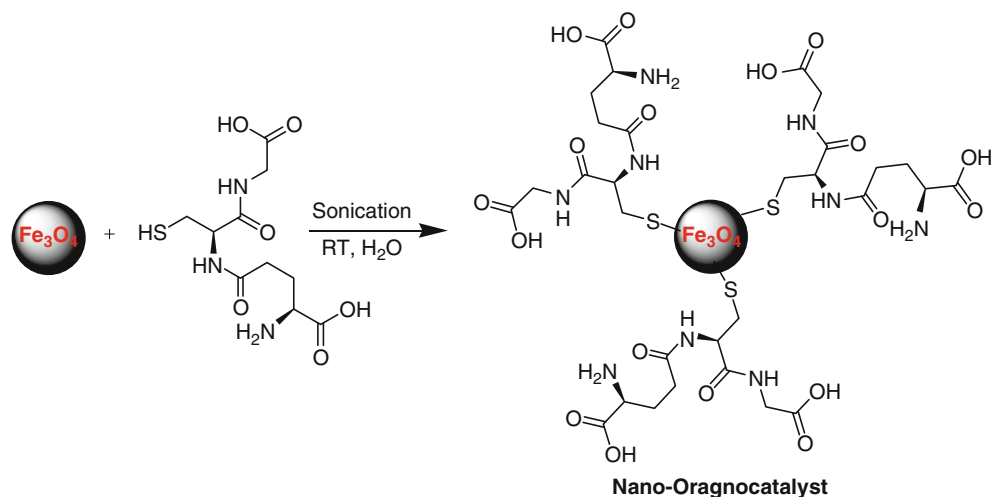
After completion of the reaction and when the stirring is stopped, the reaction mixture turns clear and catalyst is deposited on the magnetic bar because of the paramagnetic nature of the nano-Ru(OH)_x. The catalyst is conveniently removed using an external magnet, thus avoiding even a filtration step. After separation of catalyst, the clear reaction mixture is cooled slowly and crystals of benzamides with acceptable purity are precipitated. The complete operation is conducted in pure aqueous medium and no organic solvents are used during the reaction or in the workup steps.

Glutathione-Based Nano-organocatalyst for Aqueous MW-Assisted Synthesis of Heterocycles Organocatalysis has been a very active area of research during the past decade and this metal-free approach has attracted universal interest. Although a wide range of reactions has been successfully introduced using this strategy, most of these transformations are generally conducted in organic solvents. In aqueous protocols, it was observed that the addition of water often accelerated the organocatalyst-mediated reaction, making the overall protocol efficient and eco-friendly [136–138]. However, most of these methods use small amounts of water as reaction medium and excessive amounts of volatile organic solvents are used during the workup, which unfortunately defeats the central idea of



Green Chemistry with Microwave Energy. Scheme 44

Hydration of nitrile using nano-Ru(OH)_x catalyst. From Polshettiwar and Varma (Ref. [119])



Green Chemistry with Microwave Energy. Scheme 45

Nano-ferrite functionalization using glutathione. From Polshettiwar and Varma (Ref. [119])

reducing the environmental burden of organic contaminants [139].

These drawbacks were successfully circumvented in a green and sustainable manner using glutathione-based nano-organocatalyst under aqueous MW conditions [140–142]. Glutathione, a tripeptide consisting of glutamic acid, cysteine, and glycine units, is a ubiquitous antioxidant present in human and plant cells. The use of glutathione as an active catalytic moiety is preferred due to its benign nature as well as the presence of the highly active thiol group, which can be used for attachment to solid support (ferrites). The catalyst is conveniently prepared by sono-chemical covalent anchoring of glutathione molecules via coupling of its thiol group with the free hydroxyl groups of ferrite surfaces (Scheme 45).

The successful use of this glutathione-based nano-organocatalyst approach was demonstrated by Varma et al. for the synthesis of a series of pyrrole heterocycles by Paal–Knorr reaction under aqueous MW conditions. It showed excellent catalytic activity and several amines reacted with tetrahydro-2,5-dimethoxyfuran to produce the respective pyrrole derivatives in good yields (Scheme 46) [140–142].

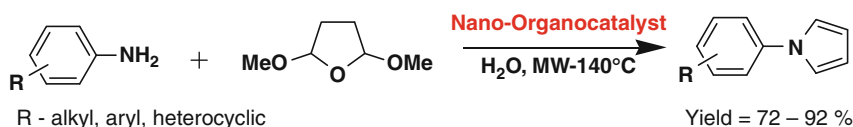
Using this strategy, various hydrazines and hydrazides were found to react efficiently with 1,3-diketones thus affording the desired pyrazoles in good yields (Scheme 47) [140–141]. All these reactions proceed

efficiently in aqueous medium and get completed in less than 20 min under MW irradiation conditions. This general approach has recently been extended to the homocoupling of boronic acids in aqueous medium under the influence of microwaves [142].

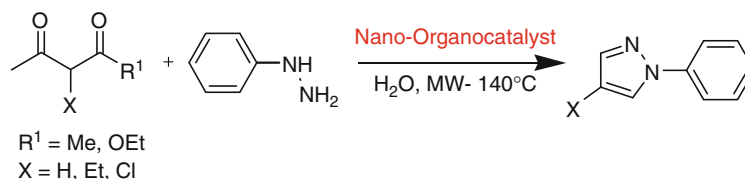
Separation of the catalyst and final product from the reaction mixture is one of the most vital aspects of synthetic protocols. Catalyst recovery occurs generally via filtration or extractive isolation of products, both of which require excessive amounts of organic solvents. However, in the aforementioned protocols, within a few seconds after stirring is stopped, catalyst gets deposited on the magnetic bar and can be easily removed using an external magnet. In most of the experiments, after completion of the reactions, the phase separation of the desired product from the aqueous medium occurs; this facilitates the isolation of synthesized heterocycles by simple decantation, without using any volatile organic solvents during the reaction or during product workup (Schemes 46, 47).

Future Directions

The use of various alternative pathways in green chemistry domain such as solvent-free synthesis and transformations [143–145], mechanochemical reactions by grinding [146–149], reactions in benign solvents like

**Green Chemistry with Microwave Energy. Scheme 46**

Paal–Knorr reactions using nano-organocatalyst. From Polshettiwar and Varma (Ref. [119])

**Green Chemistry with Microwave Energy. Scheme 47**

Synthesis of pyrazoles using nano-organocatalyst. From Polshettiwar and Varma (Ref. [119])

water [150–153], polyethylene glycol (PEG) [154, 155], and eco-friendly generated ionic liquids [44–48] and their utility in catalysis [156–161] can be augmented nicely by the use of microwave irradiation heating technique.

The “greener” production of nanoparticles [162] with relatively attractive toxicological profile [163, 164] and their enhanced utility in catalysis [115, 116, 135, 140–142] and environmental remediation [165–167] is garnering attention.

Nano-catalysts mimic homogeneous (easily accessible, high surface area) as well as heterogeneous (stable, easy to handle and isolate) catalyst systems. Nano-catalysts possessing a paramagnetic core thus allow rapid and selective chemical transformations with excellent product yield coupled with the ease of catalyst separation and recovery. Among these options, nanocatalyst-catalyzed transformations in aqueous reaction medium are one of the ideal solutions for the development of green and sustainable protocols. However, execution of many organic reactions in water is not simple due to the inherent limitation of solubility of nonpolar reactants in polar aqueous medium, which can be assisted by using MW irradiation conditions. Thus, a combination of benign water medium, nonconventional MW heating, and nano-catalyst seems to be the optimum pathway to develop the next generation of highly efficient processes [120].

Disclaimer

The views expressed in this article are those of the author and do not necessarily reflect the views and policies of the US Environmental Protection Agency. The use of trade names does not imply endorsement by the US Government.

Bibliography**Primary Literature**

1. Anastas PT, Warner JC (2000) Green chemistry: theory and practice. Oxford University Press, Oxford
2. Varma RS (1999) Solvent-free organic syntheses using supported reagents and microwave irradiation. *Green Chem* 1:43–55
3. Varma RS (1999) Solvent-free syntheses of heterocycles using microwave irradiation. *J Heterocyclic Chem* 36:1565–1571
4. Varma RS (2000) Clay and clay-supported reagents in organic synthesis. *Tetrahedron* 58:1235–1255
5. Polshettiwar V, Varma RS (2008) Microwave-assisted organic synthesis and transformations using benign reaction media. *Acc Chem Res* 41:629–639
6. Polshettiwar V, Varma RS (2008) Aqueous microwave chemistry: a clean and green synthetic tool for rapid drug discovery. *Chem Soc Rev* 37:1546–1557
7. Li C-J, Chen L (2006) Organic chemistry in water. *Chem Soc Rev* 35:68–82
8. Varma RS (2007) Clean chemical synthesis in water. *Org Chem Highlight*. <http://www.organic-chemistry.org/Highlights/2007/01February.shtm>

9. Strauss CR, Trainor RW (1995) Developments in microwave-assisted organic chemistry. *Aust J Chem* 48:1665–1692
10. Anonymous (1924) Using chemical reagents on porous carriers. *Akt –Ges Fur Chemiewerte Brit Pat* 231: 901 [Chem Abst (1925) 19: 3571]
11. Laszlo P (1987) Preparative chemistry using supported reagents. Academic, San Diego
12. Smith K (1992) Solid supports and catalyst in organic synthesis. Ellis Horwood, Chichester
13. Clark JH (1994) Catalysis of organic reactions by supported inorganic reagents. VCH, New York
14. McKillop A, Young KW (1979) Organic synthesis using supported reagents – Part I & Part II. *Synthesis* 401–422 and 481–500
15. Cornelis A, Laszlo P (1985) Clay-supported copper(II) and iron (III) nitrates: novel multi-purpose reagents for organic synthesis. *Synthesis* 100:909–918
16. Gedye R, Smith F, Westaway K, Humera A, Baldisera L, Laberge L, Rousell J (1986) The use of microwave ovens for rapid organic synthesis. *Tetrahedron Lett* 27:279–282
17. Giguere RJ, Bray TL, Duncan SM, Majetich G (1986) Application of commercial microwave ovens to organic synthesis. *Tetrahedron Lett* 27:4945–4948
18. Varma RS (2002) Advances in green chemistry: chemical syntheses using microwave irradiation. AstraZeneca Research Foundation India, Bangalore [85 Reaction schemes, ~300 references]
19. Varma RS (2002) Organic synthesis using microwaves and supported reagents. In: Loupy A (ed) *Microwaves in organic synthesis*, Chapter 6. Wiley-VCH, New York, pp 181–218
20. Pillai UR, Sahle-Demessie E, Varma RS (2002) Environmentally friendlier organic transformations on mineral supports under non-traditional conditions. *J Mater Chem* 12:3199–3207
21. Varma RS (2001) Solvent-free accelerated organic syntheses using microwaves. *Pure Appl Chem* 73:193–198
22. Perreux L, Loupy A (2001) A tentative rationalization of microwave effects in organic synthesis according to the reaction medium, and mechanistic considerations. *Tetrahedron* 57:9199–9223
23. Loupy A, Petit A, Hamelin J, Texier-Boullet F, Jacquault P, Mathe D (1998) New solvent-free organic synthesis using focused microwaves. *Synthesis* 1998:1213–1234
24. Gutierrez E, Loupy A, Bram G, Ruiz-Hitzky E (1989) Inorganic solids in “dry media” an efficient way for developing microwave irradiation activated organic reactions. *Tetrahedron Lett* 30:945–948
25. Greene TW, Wuts PGM (1991) Protective groups in organic synthesis, 2nd edn. Wiley, New York
26. Varma RS, Varma M, Chatterjee AK (1993) Microwave-assisted deacetylation on alumina: a simple deprotection method. *J Chem Soc Perkin Trans –1* 999–1000
27. Varma RS, Chatterjee AK, Varma M (1993) Alumina-mediated deacetylation of benzaldehyde diacetates. A simple deprotection method. *Tetrahedron Lett* 34:3207–3210
28. Varma RS, Chatterjee AK, Varma M (1993) Alumina-mediated microwave thermolysis: a new approach to deprotection of benzyl esters. *Tetrahedron Lett* 34:4603–4606
29. Varma RS, Lamture JB, Varma M (1993) Alumina-mediated cleavage of *t*-butyldimethylsilyl ethers. *Tetrahedron Lett* 34:3029–3032
30. Varma RS, Kumar D (1999) Microwave-accelerated solvent-free synthesis of thioketones, thiolactones, thioamides thionoesters and thioflavonoids. *Org Lett* 1:697–700
31. Varma RS, Saini RK (1997) Solid state dethioacetalization using clayfen. *Tetrahedron Lett* 38:2623–2624
32. Varma RS, Saini RK, Dahiya R (1997) Active manganese dioxide on silica: oxidation of alcohols under solvent-free conditions using microwaves. *Tetrahedron Lett* 38:7823–7824
33. Namboodiri VV, Polshettiwar V, Varma RS (2007) Expeditious oxidation of alcohols to carbonyl compounds using iron (III) nitrate. *Tetrahedron Lett* 48:8839–8842
34. Varma RS, Dahiya R, Saini RK (1997) Iodobenzene diacetate on alumina: rapid oxidation of alcohols to carbonyl compounds in solventless system using microwaves. *Tetrahedron Lett* 38:7029–7032
35. Varma RS, Dahiya R, Kumar D (1998) Solvent-free oxidation of benzoin using oxone[®] on wet alumina under microwave irradiation. *Molecules Online* 2:82–85
36. Varma RS, Saini RK, Meshram HM (1997) Selective oxidation of sulfides to sulfoxides and sulfones by microwave thermolysis on wet silica-supported sodium periodate. *Tetrahedron Lett* 38:6525–6528
37. Varma RS, Kumar D (1999) Solid state oxidation of 1,4-dihydropyridines to pyridines using phenyliodine(III) bis (trifluoroacetate) or elemental sulfur. *J Chem Soc Perkin Trans –1* 1755–1757
38. Varma RS, Saini RK (1997) Microwave-assisted reduction of carbonyl compounds in solid state using sodium borohydride supported on alumina. *Tetrahedron Lett* 38:4337–4338
39. Varma RS, Dahiya R (1998) Sodium borohydride on wet clay: solvent-free reductive amination of carbonyl compounds using microwaves. *Tetrahedron* 54:6293–6298
40. Vass A, Dudas J, Toth J, Varma RS (2001) Solvent-free reduction of aromatic nitro compounds with alumina-supported hydrazine under microwave irradiation. *Tetrahedron Lett* 42:5347–5349
41. Varma RS, Naicker KP (1998) Hydroxylamine on clay: a direct synthesis of nitriles from aromatic aldehydes using microwaves under solvent-free conditions. *Molecules Online* 2: 94–96
42. Welton T (2004) Ionic liquids in catalysis. *Coord Chem Rev* 248:2459–2477
43. Rogers RD, Seddon KN, Volkove S (2002) Green Industrial applications of ionic liquids. *NOTO science, series*
44. Varma RS, Namboodiri VV (2001) An expeditious solvent-free route to ionic liquids using microwaves. *Chem. Commun* 643–644
45. Namboodiri VV, Varma RS (2002) An improved preparation of 1, 3-dialkylimidazolium tetrafluoroborate ionic liquids using microwaves. *Tetrahedron Lett* 43:5381–5383
46. Namboodiri VV, Varma RS (2002) Solvent-free sonochemical preparation of ionic liquids. *Org Lett* 4:3161–3163

47. Kim YJ, Varma RS (2005) Microwave-assisted preparation of imidazolium-based tetrachloroindate (III) and their application in the tetrahydropyranylation of alcohols. *Tetrahedron Lett* 46:1467–1469
48. Kim YJ, Varma RS (2005) Microwave-assisted preparation of 1-butyl-3-methylimidazolium tetrachlorogallate and its catalytic use in acetal formation under mild conditions. *Tetrahedron Lett* 46:7447–7449
49. Swatoski RP, Spear SK, Holbrey JD, Rogers RD (2002) Dissolution of cellulose with ionic liquids. *J Am Chem Soc* 124:4974–4975
50. Varma RS, Nambodiri VV (2002) Microwave-assisted preparation of dialkylimidazolium tetrachloroaluminates and their use as catalysts in the solvent-free tetrahydropyranylation of alcohols and phenols. *Chem Commun* 342–343
51. Kim YJ, Varma RS (2005) Tetrahaloindate(III)-based ionic liquids in the coupling reaction of carbon dioxide and epoxides to generate cyclic carbonates: H-bonding and mechanistic studies. *J Org Chem* 70:7882–7891
52. Plechkova NV, Seddon KR (2008) Applications of ionic liquids in the chemical industry. *Chem Soc Rev* 37:123–150
53. Herrero MA, Kremsner JM, Kappe CO (2008) Nonthermal microwave effects revisited – on the importance of internal temperature monitoring and agitation in microwave chemistry. *J Org Chem* 73:36–47
54. Polshettiwar V, Varma RS (2007) Greener and sustainable approaches to the synthesis of pharmaceutically active heterocycles. *Curr Opin Drug Discov Devel* 10:723–737
55. Chen J, Spear SK, Huddleston JG, Rogers RD (2005) Polyethylene glycol and solutions of polyethylene glycol as green reaction media. *Green Chem* 7:64–82
56. Leadbeater NE, Marco M (2003) Rapid and amenable Suzuki coupling reaction in water using microwave and conventional heating. *J Org Chem* 68:888–892
57. Crozet MD, Castera-Ducros C, Vanelle P (2006) An efficient microwave-assisted Suzuki cross-coupling reaction of imidazo [1, 2-a] pyridines in aqueous medium. *Tetrahedron Lett* 47:7061–7065
58. Zhu R, Qu F, Quelever G, Peng L (2007) Direct synthesis of 5-aryltriazole acyclonucleosides via Suzuki coupling in aqueous solution. *Tetrahedron Lett* 48:2389–2393
59. Dawood KM (2007) Microwave-assisted Suzuki–Miyaura and Heck–Mizoroki cross-coupling reactions of aryl chlorides and bromides in water using stable benzothiazole-based palladium (II) precatalysts. *Tetrahedron* 63:9642–9651
60. Arvela RK, Leadbeater NE (2005) Microwave-promoted Heck coupling using ultralow metal catalyst concentrations. *J Org Chem* 70:1786–1790
61. Arvela RK, Pasquini S, Larhed M (2007) Highly regioselective internal Heck arylation of hydroxyalkyl vinyl ethers by aryl halides in neat water. *J Org Chem* 72:6390–6396
62. Appukkuttan P, Dehaen W, der Eycken EV (2003) Transition-metal-free Sonogashira-type coupling reactions in water. *Eur J Org Chem* 2003:4713–4716
63. Alcida E, Najera C (2006) The first fluoride-free Hiyama reaction of vinylsiloxanes promoted by sodium hydroxide in water. *Adv Synth Catal* 348:2085–2091
64. Kaval N, Bisztray K, Dehaen W, Kappe CO, der Eycken EV (2003) Microwave-enhanced transition metal-catalyzed decoration of 2(1H)-pyrazinone scaffolds. *Mol Divers* 7:125–133
65. Miyazawa A, Tanaka K, Sakakura T, Tashiro M, Tashiro H, Surya Prakash GK, Olah GA (2005) Microwave-assisted direct transformation of amines to ketones using water as an oxygen source. *Chem Commun* 2104–2106
66. Kumar V, Sharma A, Sharma A, Sinha AK (2007) Remarkable synergism in methylimidazole-promoted decarboxylation of substituted cinnamic acid derivatives in basic water medium under microwave irradiation: a clean synthesis of hydroxylated (E)-stilbenes. *Tetrahedron* 63:7640–7646
67. Dallinger D, Kappe CO (2007) Microwave-assisted synthesis in water as solvent. *Chem Rev* 107:2563–2591
68. Garuti L, Roberti M, Pizzirani D (2007) Nitrogen-containing heterocyclic quinones: a class of potential selective antitumor agents. *Mini Rev Med Chem* 7:481–489
69. Sperry JB, Wright DL, Furans (2005) Thiophenes and related heterocycles in drug discovery. *Curr Opin Drug Discov Devel* 8:723–740
70. Kappe CO (2002) High-speed combinatorial synthetics utilizing microwave irradiation. *Curr Opin Chem Biol* 6:314–320
71. Polshettiwar V, Varma RS (2008) Greener and expeditious synthesis of bio-active heterocycles using microwave irradiation. *Pure Appl Chem* 80:777–790
72. Roberts BA, Strauss CR (2005) Toward rapid, “green”, predictable microwave-assisted synthesis. *Acc Chem Res* 38:653–661
73. Kappe CO (2004) Controlled microwave heating in modern organic synthesis. *Angew Chem Int Ed* 43:6250–6284
74. Ju Y, Varma RS (2004) Aqueous *N*-alkylation of amines using alkyl halides: direct generation of tertiary amines under microwave irradiation. *Green Chem* 6:219–221
75. Ju Y, Varma RS (2005) An efficient and simple aqueous *N*-heterocyclization of aniline derivatives: microwave-assisted synthesis of *N*-aryl azacycloalkanes. *Org Lett* 7:2409–2411
76. Ju Y, Varma RS (2005) Microwave-assisted cyclocondensation of hydrazine derivatives with alkyl dihalides or ditosylates in aqueous media: syntheses of pyrazole, pyrazolidine and phthalazine derivatives. *Tetrahedron Lett* 46:6011–6014
77. Ju Y, Varma RS (2006) Aqueous *N*-heterocyclization of primary amines and hydrazines with dihalides: microwave-assisted syntheses of *N*-azacycloalkanes, isoindole, pyrazole, pyrazolidine, and phthalazine derivatives. *J Org Chem* 71:135–141
78. Ju Y, Li C-J, Varma RS (2004) Microwave-assisted Cu (I) catalyzed solvent-free three component coupling of aldehyde, alkyne and amine. *QSAR Comb Sci* 23:891–894
79. Kim YJ, Varma RS (2004) Microwave-assisted preparation of cyclic ureas from diamines in the presence of ZnO. *Tetrahedron Lett* 45:7205–7208
80. Varma RS, Kumar D (1999) Microwave-accelerated three-component condensation reaction on clay: solvent-free synthesis of imidazo [1, 2-a] annulated pyridines, pyrazines and pyrimidones. *Tetrahedron Lett* 40:7665–7669
81. Kappe CO, Kumar D, Varma RS (1999) Microwave-assisted high-speed parallel synthesis of 4-aryl-3,

- 4-dihydropyrimidin-2(1H)-ones using a solventless Biginelli condensation protocol. *Synthesis* 10:1799–1803
82. Polshettiwar V, Varma RS (2007) Biginelli reaction in aqueous medium: a greener and sustainable approach to substituted 3, 4-dihydropyrimidin-2(1H)-ones. *Tetrahedron Lett* 48: 7343–7346
83. Varma RS, Dahiya R (1998) An expeditious and solvent-free synthesis of 2-amino-substituted isoflav-3-enes using microwave irradiation. *J Org Chem* 63:8038–8041
84. Varma RS, Kumar D, Liesen PJ (1998) Solid state synthesis of 2-arylbenzo[b]furans, 1,3-thiazoles and 3-aryl-5,6-dihydroimidazo [2,1-b][1,3] thiazoles from α -tosyloxyketones using microwave irradiation. *J Chem Soc Perkin Trans* —1 4093–4096
85. Polshettiwar V, Varma RS (2007) Tandem bis-aldol reaction of ketones: a facile one pot synthesis of 1, 3-dioxanes in aqueous medium. *J Org Chem* 72:7420–7422
86. Jeselnik M, Varma RS, Polanc S, Kocivar M (2001) Catalyst-free reactions under solvent-free conditions: microwave-assisted synthesis of heterocyclic hydrazones below the melting points of neat reactants. *Chem Commun* 1716–1717
87. Polshettiwar V, Varma RS (2007) Polystyrene sulfonic acid catalyzed greener synthesis of hydrazones in aqueous medium using microwaves. *Tetrahedron Lett* 48:5649–5652
88. Polshettiwar V, Varma RS (2008) Rapid access to bio-active heterocycles: one-pot solvent-free synthesis of 1, 3, 4-oxadiazoles and 1, 3, 4-thiadiazoles. *Tetrahedron Lett* 49:879–883
89. Nadagouda MN, Varma RS (2007) Microwave-assisted shape-controlled bulk synthesis of noble nanocrystals and their catalytic properties. *Cryst Growth Des* 7:686–690
90. Baruwati B, Varma RS (2009) High value products from waste: grape pomace extract – a three-in-one package for the synthesis of metal nanoparticles. *ChemSusChem* 2:1041–1044
91. Baruwati B, Polshettiwar V, Varma RS (2009) Glutathione promoted expeditious green synthesis of silver nanoparticles in water using microwaves. *Green Chem* 11:926–930
92. Nadagouda MN, Polshettiwar V, Varma RS (2009) Self-assembly of palladium nanoparticles: synthesis of nanobelts, nanoplates and nanotrees using vitamin B₁ and their application in carbon-carbon coupling reactions. *J Mater Chem* 19:2026–2031
93. Nadagouda MN, Varma RS (2006) Green and controlled synthesis of gold and platinum nanomaterials using vitamin B₂: density-assisted self-assembly of nanospheres, wires and rods. *Green Chem* 8:516–518
94. Nadagouda MN, Varma RS (2007) A greener synthesis of core (Fe, Cu)-shell (Au, Pt, Pd and Ag) nanocrystals using aqueous vitamin C. *Cryst Growth Des* 7:2582–2587
95. Nadagouda MN, Varma RS (2008) Green synthesis of silver and palladium nanoparticles at room temperature using coffee and tea extract. *Green Chem* 10:859–862
96. Nadagouda MN, Varma RS (2008) Microwave-assisted shape controlled bulk synthesis of Ag and Fe nanorods in poly (ethylene glycol) solutions. *Cryst Growth Des* 8:291–295
97. Nadagouda MN, Varma RS (2007) Synthesis of thermally stable carboxymethyl cellulose/metal biodegradable nanocomposite films for potential biological applications. *Biomacromolecules* 8:2762–2767
98. Nadagouda MN, Varma RS (2007) Microwave-assisted synthesis of cross-Linked poly (vinyl alcohol) nanocomposites comprising single-wall carbon nanotubes (SWNT), multi-wall carbon nanotubes (MWNT) and buckminsterfullerene (C-60). *Macromol Rapid Commun* 28:842–847
99. Polshettiwar V, Nadagouda MN, Varma RS (2007) Synthesis and applications of micro-pine structured nano-catalyst. *Chem Commun* 6318–6320
100. Polshettiwar V, Baruwati B, Varma RS (2009) Self-assembly of metal oxides into three-dimensional nanostructures: synthesis and application in catalysis. *ACS Nano* 3:728–736
101. Mamedov AA, Belov A, Giersig M, Mamedova NN, Kotov NA (2001) Nanorainbows: graded semiconductor films from quantum dots. *J Am Chem Soc* 123:7738–7739
102. Alivisatos P (1996) Semiconductor clusters, nanocrystals, and quantum dots. *Science* 271:933–937
103. Klimov VI, Mikhailovsky AA, Xu S, Malko A, Hollingsworth JA, Leatherdale CA, Eisler H-J, Bawendi MG (2000) Optical gain and stimulated emission in nanocrystal quantum dots. *Science* 290:314–317
104. Sundar VC, Eisler H-J, Bawendi MG (2002) Room-temperature, tunable gain media from novel II-VI nanocrystal-titania composite matrices. *Adv Mater* 14:739–743
105. Bruchez M, Moronne M, Gin P, Weiss S, Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science* 281:2013–2016
106. Chan WCW, Nie S (1998) Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science* 281:2016–2018
107. Michalet X, Pinaud FF, Bentolila LA, Tsay JM, Doose S, Li JJ, Sundaresan G, Wu AM, Gambhir SS, Weiss S (2005) Quantum dots for live cells, in vivo imaging, and diagnostics. *Science* 307:538–544
108. Murray CB, Norris DJ, Bawendi MG (1993) Synthesis and characterization of nearly monodisperse CdE (E=sulfur, selenium, tellurium) semiconductor nanocrystallites. *J Am Chem Soc* 115:8706–8715
109. Qian H, Qiu X, Li L, Ren J (2006) Microwave-assisted aqueous synthesis: a rapid approach to prepare highly luminescent ZnSe(S) alloyed quantum dots. *J Phys Chem B* 110:9034–9040
110. Zhao Y, Hong J-M, Zhu J-J (2004) Microwave-assisted self-assembled ZnS nanoballs. *J Cryst Growth* 270:438–445
111. Zhu J, Palchik O, Chen S, Gedanken A (2000) Microwave assisted preparation of CdSe, PbSe, and Cu_{2-x}Se nanoparticles. *J Phys Chem B* 104:7344–7347
112. Schumacher W, Nagy A, Waldman WJ, Dutta PK (2009) Direct synthesis of aqueous CdSe/ZnS-based quantum dots using microwave irradiation. *J Phys Chem C* 113:12132–12139
113. Qian H, Li L, Ren J (2005) One-step and rapid synthesis of high quality alloyed quantum dots (CdSe–CdS) in aqueous phase by microwave irradiation with controllable temperature. *Mater Res Bull* 40:1726–1736

114. Lu A-H, Salabas EL, Schuth F (2007) Magnetic nanoparticles: synthesis, protection, functionalization, and application. *Angew Chem Int Ed* 46:1222–1244
115. Polshettiwar V, Varma RS (2009) Nanoparticle-supported and magnetically recoverable palladium (Pd) catalyst: a selective and sustainable oxidation protocol with high turnover number. *Org Biomol Chem* 7:37–40
116. Polshettiwar V, Baruwati B, Varma RS (2009) Nanoparticle-supported and magnetically recoverable nickel catalyst: a robust and economic hydrogenation and transfer hydrogenation protocol. *Green Chem* 11:127–131
117. Polshettiwar V, Nadagouda MN, Varma RS (2008) Synthesis and applications of micro-pine structured nano-catalyst. *Chem. Commun* 6318–6320
118. Baruwati B, Guin D, Manorama SV (2007) Pd on surface-modified NiFe_2O_4 nanoparticles: a magnetically recoverable catalyst for Suzuki and Heck reactions. *Org Lett* 9:5377–5380
119. Guin D, Baruwati B, Manorama SV (2007) Pd on amine-terminated ferrite nanoparticles: a complete magnetically recoverable facile catalyst for hydrogenation reactions. *Org Lett* 9:1419–1421
120. Polshettiwar V, Varma RS (2010) Green chemistry by nanocatalysis. *Green Chem* 12:743–754
121. Loupy A, Varma RS (2006) Microwave effects in organic synthesis: mechanistic and reaction medium considerations. *Chim Oggi* 24:36–40
122. Strauss CR, Varma RS (2006) Microwaves in green and sustainable chemistry. *Top Curr Chem* 266:199–231
123. Kappe CO, Dallinger D (2009) Controlled microwave heating in modern organic synthesis: highlights from the 2004–2008 literature. *Mol Divers* 13:71–193
124. Polshettiwar V, Nadagouda MN, Varma RS (2009) Microwave-assisted chemistry: a rapid and sustainable route to synthesis of organics and nanomaterials. *Aust J Chem* 62:16–26
125. Gabriel C, Gabriel S, Grant EH, Halstead BSJ, Mingos DMP (1998) Dielectric parameters relevant to microwave dielectric heating. *Chem Soc Rev* 27:213–224
126. Poliokoff M, Licence P (2007) Sustainable technology: green chemistry. *Nature* 450:810–812
127. Polshettiwar V, Varma RS (2008) Olefin ring closing metathesis and hydrosilylation reaction in aqueous medium by Grubbs second generation ruthenium catalyst. *J Org Chem* 73:7417–7419
128. Hoz A, Diaz-Ortiz A, Moreno A (2005) Microwaves in organic synthesis. Thermal and non-thermal microwave effects. *Chem Soc Rev* 34:164–178
129. Nilsson P, Larhed M, Hallberg A (2001) Highly regioselective, sequential, and multiple palladium-catalyzed arylations of vinyl ethers carrying a coordinating auxiliary: an example of a Heck triarylation process. *J Am Chem Soc* 123:8217–8225
130. Kaiser NFK, Bremberg U, Larhed M, Moberg C, Hallberg A (2000) Fast, convenient, and efficient molybdenum-catalyzed asymmetric allylic alkylation under noninert conditions: an example of microwave-promoted fast chemistry. *Angew Chem Int Ed* 39:3596–3598
131. Bogdal D, Lukasiewicz M, Pielichowski J, Miciak A, Sz B (2003) Microwave-assisted oxidation of alcohols using Magtrieve™. *Tetrahedron* 59:649–653
132. Razzak T, Kremser JM, Kappe CO (2008) Investigating the existence of nonthermal/specific microwave effects using silicon carbide heating elements as power modulators. *J Org Chem* 73:6321–6329
133. Leadbeater NE, Torrenius HM (2002) A study of the ionic liquid mediated microwave heating of organic solvents. *J Org Chem* 67:3145–3148
134. Cadierno V, Francos J, Gimeno J (2008) Selective ruthenium-catalyzed hydration of nitriles to amides in pure aqueous medium under neutral conditions. *Chem Eur J* 14:6601–6605
135. Polshettiwar V, Varma RS (2009) Nanoparticle-supported and magnetically recoverable ruthenium hydroxide catalyst: efficient hydration of nitriles to amides in aqueous medium. *Chem Eur J* 15:1582–1586
136. Brogan AP, Dickerson TJ, Janda KD (2006) Enamine-based aldol organocatalysis in water: are they really all wet? *Angew Chem Int Ed* 45:8100–8102
137. Hayashi Y, Samanta S, Gotoh H, Ishikawa H (2008) Asymmetric Diels-Alder reactions of,unsaturated aldehydes catalyzed by a diarylprolinol silyl ether salt in the presence of water. *Angew Chem Int Ed* 47:6634–6637
138. Huang J, Zhang X, Armstrong DW (2007) Highly efficient asymmetric direct stoichiometric aldol reactions on/in water. *Angew Chem Int Ed* 46:9073–9077
139. Blackmond DG, Armstrong A, Coombe V, Wells A (2007) Water in organocatalytic processes: debunking the myths. *Angew Chem Int Ed* 46:3798–3800
140. Polshettiwar V, Baruwati B, Varma RS (2009) Magnetic nanoparticle-supported glutathione: a conceptually sustainable organocatalyst. *Chem Commun* 1837–1839
141. Polshettiwar V, Varma RS (2010) Nano-organocatalyst: magnetically retrievable ferrite-anchored glutathione for microwave-assisted Paal-Knorr reaction, aza-Michael addition and pyrazole synthesis. *Tetrahedron* 66:1091–1097
142. Luque R, Baruwati B, Varma RS (2010) Magnetically separable nanoferrite-anchored glutathione: aqueous homocoupling of arylboronic acids under microwave irradiation. *Green Chem* 12:1540–1543. doi:10.1039/C0GC00083C
143. Polshettiwar V, Varma RS (2008) Ring-fused amins: catalyst and solvent-free microwave-assisted α -amination of nitrogen heterocycles. *Tetrahedron Lett* 49:7165–7167
144. Varma RS, Naicker KP, Liesen PJ (1998) Microwave-accelerated crossed Cannizzaro reaction using barium hydroxide under solvent-free conditions. *Tetrahedron Lett* 3:8437–8440
145. Pillai UR, Sahle-Demessie E, Nambodiri VV, Varma RS (2002) An efficient and ecofriendly oxidation of alkenes using iron nitrate and molecular oxygen. *Green Chem* 4:495–497
146. Kumar D, Chandra Sekhar KVG, Dhillion H, Rao VS, Varma RS (2004) An expeditious synthesis of 1-aryl-4-methyl-1, 2, 4-triazolo [4, 3-a] quinoxalines under solvent-free conditions using iodobenzene diacetate. *Green Chem* 6:156–157

147. Kumar D, Sundaree MS, Patel G, Rao VS, Varma RS (2006) Solvent-free facile synthesis of novel α -tosyloxy β -keto sulfones using [hydroxy(tosyloxy)iodo] benzene. *Tetrahedron Lett* 47:8239–8241
148. Kumar D, Sundaree MS, Rao VS, Varma RS (2006) A facile one-pot synthesis of β -keto sulfones from ketones under solvent-free conditions. *Tetrahedron Lett* 47:4197–4199
149. Varma RS (2008) Chemical activation by mechanochemical mixing, microwave, and ultrasonic irradiation. *Green Chem* 10:1129–1130
150. Polshettiwar V, Varma RS (2007) Tandem bis-aza-Michael addition reaction of amines in aqueous medium promoted by polystyrenesulfonic acid. *Tetrahedron Lett* 48:8735–8738
151. Kumar D, Reddy VB, Mishra BG, Rana RK, Nadagouda MN, Varma RS (2007) Nanosized magnesium oxide as catalyst for the rapid and green synthesis of substituted 2-amino-2-chromenes. *Tetrahedron* 63:3093–3097
152. Skouta R, Varma RS, Li CJ (2005) Efficient Trost's γ -addition catalyzed by reusable polymer-supported triphenylphosphine in aqueous media. *Green Chem* 7:571–575
153. Ju Y, Kumar D, Varma RS (2006) Revisiting nucleophilic substitution reactions: microwave-assisted synthesis of azides, thiocyanates, and sulfones in an aqueous medium. *J Org Chem* 71:6697–6700
154. Namboodiri VV, Varma RS (2001) Microwave-accelerated Suzuki cross-coupling reaction in polyethylene glycol (PEG). *Green Chem* 3:146–148
155. Kumar D, Patel G, Mishra BG, Varma RS (2008) Ecofriendly polyethylene glycol (PEG)-promoted Michael addition reactions of α , β -unsaturated compounds. *Tetrahedron Lett* 49:6974–6976
156. Keh CCK, Namboodiri VV, Varma RS, Li C-J (2002) Direct formation of tetrahydropyrans via catalysis in ionic liquid. *Tetrahedron Lett* 43:4993–4996
157. Li Z, Wei C, Varma RS, Li C-J (2004) Three-component coupling of aldehyde, alkyne, and amine catalyzed by silver in ionic liquid. *Tetrahedron Lett* 45:2443–2446
158. Yang X-F, Wang M, Varma RS, Li C-J (2003) Aldol- and Mannich-type reactions via in situ olefin migration in ionic liquid. *Org Lett* 5:657–660
159. Yang X-F, Wang M, Varma RS, Li C-J (2004) Ruthenium-catalyzed tandem olefin migration aldol and Mannich-type reactions in ionic liquid. *J Mol Catal A Chem* 214:147–154
160. Yoo K, Namboodiri VV, Varma RS, Smirniotis PG (2004) Ionic liquid-catalyzed alkylation of isobutane with 2-butene. *J Catal* 222:511–519
161. Namboodiri VV, Varma RS, Sahle-Demessie E, Pillai UR (2002) Selective oxidation of styrene to acetophenone in the presence of ionic liquids. *Green Chem* 4:170–173
162. Nadagouda MN, Hoag GE, Collins JB, Varma RS (2009) Green synthesis of Au nanostructures at room temperature using biodegradable plant surfactants. *Cryst Growth Des* 9:4979–4983
163. Nadagouda MN, Castle A, Murdock RC, Hussain SM, Varma RS (2010) In vitro biocompatibility of nanoscale zerovalent iron particles (nZVI) synthesized using tea polyphenols. *Green Chem* 12:114–122
164. Moulton MC, Braydich-Stolle LK, Nadagouda MN, Kunzelman S, Hussain SM, Varma RS (2010) Synthesis, characterization and biocompatibility of "green" synthesized silver nanoparticles using tea polyphenols. *Nanoscale* 2:763–770
165. Hoag GE, Collins JB, Holcomb JL, Hoag JR, Nadagouda MN, Varma RS (2009) Degradation of bromothymol blue by 'greener' nano-scale zerovalent iron synthesized using tea polyphenols. *J Mater Chem* 19:8671–8677
166. Virkutyte J, Varma RS (2010) Fabrication and visible-light photocatalytic activity of novel Ag/TiO_{2-x}N_x photocatalyst. *New J Chem* 34:1094–1096
167. Virkutyte J, Baruwati B, Varma RS (2010) Visible light induced photobleaching of methylene blue over melamine doped TiO₂ nanocatalyst. *Nanoscale* 2(7):1109–1111

Books and Reviews

- Ahluwalia VK, Varma RS (2008) Alternative energy processes in chemical synthesis microwave, ultrasound and photo activation. Narosa Publishing House, New Delhi. ISBN 978-81-7319-848-9
- Ahluwalia VK, Varma RS (2009) Green solvents for organic synthesis. Narosa Publishing House, New Delhi. ISBN 978-81-7319-964-6
- Clark JH, Macquarrie D (2002) Handbook of green chemistry and technology. Blackwell Science, Oxford
- Kappe CO, Stadler A (2005) Microwaves in organic and medicinal chemistry. Wiley-VCH, Weinheim, p 410
- Kappe CO, Dallinger D, Murphree SS (2009) Practical microwave synthesis for organic chemists – strategies, instruments, and protocols. Wiley-VCH, Weinheim, p 296
- Matlack AS (2001) Introduction to green chemistry. Marcel Dekkers, New York
- Nadagouda MN, Varma RS (2009) Risk reduction via greener synthesis of noble metal nanostructures and nanocomposites. In: Linkov I, Steevens J (eds) Nanomaterials: risks and benefits-proceedings of the NATO advanced workshop. Springer, Faro, pp 209–218
- Polshettiwar V, Varma RS (2009) Environmentally benign chemical synthesis via mechanochemical mixing and microwave irradiation. In: Ballini R (ed) Eco-friendly synthesis of fine chemicals, RSC green chemistry book series. RSC, Cambridge, England, pp 275–292
- Polshettiwar V, Varma RS (2009) Non-conventional energy sources for green synthesis in water (microwave, ultrasound, and photo). In: Li C-J, Anastas PT (eds) Handbook series, Handbook of green chemistry, Vol. 5: reactions in water. Wiley-VCH, Weinheim. ISBN 978-3-527-31574-1
- Polshettiwar V, Varma RS (eds) (2010) Aqueous microwave chemistry: synthesis and applications, vol 7, RSC green chemistry series. Royal Society Chemistry, Cambridge, UK
- Strauss CR, Varma RS (2006) Microwaves in green and sustainable chemistry. In: Larhed M, Olofsson K (eds) Microwave methods

in organic synthesis, vol 266, Series in topics in current chemistry. Springer, Heidelberg, pp 199–231

Varma RS (2000) Environmentally benign organic transformations using microwave irradiation under solvent-free conditions. In: Anastas PT, Tundo P (eds) *Green chemistry: challenging perspectives*. Oxford University Press, Oxford, pp 221–244

Varma RS (2000) Expeditionary solvent-free organic syntheses using microwave irradiation. In: Anastas PT, Heine L, Williamson T (eds) *Green chemical syntheses and processes*, Chapter 23, vol 767, ACS symposium series. American Chemical Society, Washington, DC, pp 292–312

Varma RS (2001) Microwave organic synthesis. In: Geller E (ed) *McGraw-Hill Yearbook of Science and Technology 2002*. McGraw-Hill, New York, pp 223–225

Varma RS (2006) Microwave technology: chemical synthesis applications. In: Seidel A (ed) *Kirk-Othmer on-line encyclopedia of chemical technology*, vol 16, 5th edn. Wiley, Hoboken, pp 538–594

Varma RS, Ju Y (2005) Microwaves in organic synthesis. In: Afonso CAM, Crespo JG (eds) *Solventless reactions (SLR)*, Chapter 2.2. Wiley-VCH, Weinheim, pp 53–87

Varma RS, Ju Y (2006) Organic synthesis using microwaves and supported reagents. In: Loupy A (ed) *Microwaves in organic synthesis*, Chapter 8, 2nd edn. Wiley-VCH, Weinheim, pp 362–415

Green Infrastructure and Climate Change

STEPHAN PAULEIT¹, OLE FRYD², ANTJE BACKHAUS²,
MARINA BERGEN JENSEN²

¹Center of Life and Food Sciences Weihenstephan,
Technical University of Munich, Freising, Germany

²Forest and Landscape Denmark, University of
Copenhagen, Frederiksberg C, Denmark

Article Outline

Glossary

Definition of the Subject

Introduction

The Role of Green Infrastructure for Adaptation of
Cities to Climate Change

The Potential of Green Infrastructure to Mitigate
Climate Change

Vulnerability of the Green Infrastructure to Climate
Change Impacts

Future Directions

Bibliography

Glossary

Adaptation (with respect to climate change) The adjusting of systems, natural or human, in response to actual or expected impacts of climate change, such as sea level rise, to reduce vulnerability or increase resilience in response to observed or expected changes in climate and associated extreme events [1, p. 869]. A distinction has been made between planned adaptation (e.g., urban planning), which is the focus of this chapter, and autonomous adaptation (e.g., by individual action such as improving housing insulation, installing air-conditioning, etc.) [2].

Ecosystem services Are “the benefits people obtain from ecosystems” [3]. “These include provisioning services such as food, water, timber, and fiber; regulating services that affect climate, floods, disease, wastes, and water quality; cultural services that provide recreational, aesthetic, and spiritual benefits; and supporting services such as soil formation, photosynthesis, and nutrient cycling” (3, Preface: V). In urban areas, ecosystem services are clearly related to land use and land cover. Therefore, spatial planning and regulations that influence the spatial pattern and intensity of land use, and in particular the provision and quality of green spaces, can have huge implications for the ecology of cities [4, 5].

Evapotranspiration The sum of evaporation of water from surfaces and the transpiration of water by plants and animals. According to US Geological Survey [6], transpiration from plants accounts for approximately 10% of air moisture. A large oak has been estimated to transpire up to 151,000 l water per year. This would account for more than 400 l/day. More modest figures are given by other sources, for example, 200 l per day for a single, fully grown beech [7]. However, generally, these figures need to be considered with great caution as they are based on rough estimates and will greatly vary between trees depending on stand and site conditions, tree condition, as well as climatic conditions.

Green infrastructure The term “Green infrastructure” was first introduced in the USA at the end of the 1990s – it has been defined as an

“interconnected network of protected land and water that supports native species, maintains natural ecological processes, sustains air and water resources and contributes to the health and quality of life for America’s communities and people” [8]. “Urban green infrastructure” is the network of green areas in cities. The term makes reference to other types of urban infrastructures (e.g., the road system). This interpretation of green infrastructure relates to a fine-scale urban application where hybrid infrastructures of green spaces and built systems are planned and designed to support multiple ecosystem services. It has been argued that planning of an urban green infrastructure should promote multifunctionality and connectivity of urban green space. It can integrate both public and private green space. It should be based on a long-term vision and a communicative and socially inclusive approach to its planning and management [9, 10].

Mitigation (with respect to climate change) Reducing greenhouse gas emissions and enhancing sinks.

Resilience Is the ability of a system to adapt and adjust to changing internal or external processes [11, 12]. Resilience is the flip side of vulnerability – a resilient system or population is not sensitive to climate variability and change and has the capacity to adapt [13].

Sustainable urban drainage systems Sequence of management practices utilizing urban green areas for storage, infiltration, evaporation, and conveyance of stormwater runoff.

Urban heat-island effect Significantly higher temperatures experienced in cities compared to the rural surroundings as a result of changed solar reflection in the built environment, less evapotranspiration, and anthropogenic heat from combustion engines, heating of buildings, and other use of energy.

Vulnerability (with respect to climate change) Is the degree to which a system is susceptible to, and unable to cope with, adverse effects of climate change, including climate variability and extremes. Vulnerability is a function of the character, magnitude, and rate of climate change and variation to which a system is exposed, its sensitivity, and its adaptive capacity [1, p. 883].

Definition of the Subject

By 2050, two-thirds of the world’s population will live in cities. Already today, cities of the developed world are a major source of greenhouse gas emissions. Therefore, cities need to make serious efforts to mitigate climate change. Urban planning can play a major role in this respect by designing compact, low footprint cities. However, climate change will also make a severe impact on cities mostly by intensification of the heat-island effect, increase of surface runoff from more frequent and intense rainstorms, and by coastal and riverine flooding. Urban planning will play an important role for development and implementation of integrated strategies for climate change mitigation and adaptation. Notably, “green infrastructure” can assist in adapting cities to climate change by reducing the urban heat-island effect and by managing stormwater runoff. Urban greening, such as planting of shade trees and roof greening, can also reduce the energy demand for house heating and cooling. Therefore, the design of the urban landscape will have a direct influence on mitigating climate change and the impact of climate change on people’s livelihoods and assets.

Introduction

Urbanization and climate change are closely related. At present, more than half of the world’s population lives in cities. Over the next 40 years, the world’s urban population is projected to increase by more than three billion people [14]. This will almost double the current urban population and lead to massive land-use changes in and around current urban settlements.

The urban built-up areas are expected to increase by 250% and cover about one million square kilometers by 2030 [15]. This corresponds to the development of more than four hectares of new urban land every minute in the next 20 years. Ninety percent of the future urban expansion is expected to take place in Asia and Africa [14]. New urbanization is primarily an issue in developing countries whereas already 70–90% of the population of the developed world lives in urban settlements. Therefore, climate change mitigation and adaptation is an equally important issue in the developed as well as the developing world.

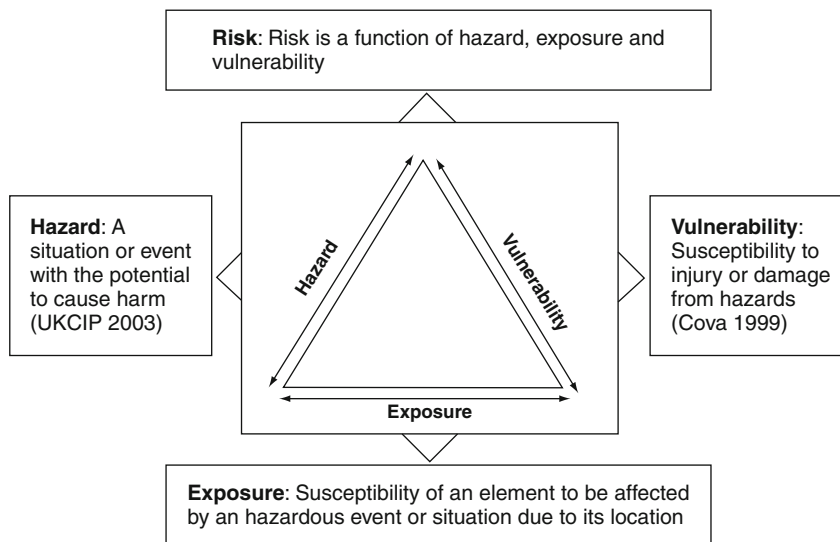
It has been estimated that 78% of carbon emissions from fossil fuel burning and cement manufacturing [16] and 85% of the anthropogenic emissions of carbon dioxide, chlorine-fluorine-carbons, and tropospheric ozone stem from urban areas [17]. While these and similar figures have been debated [18], there can be little doubt that urban areas, particularly in the developed world and in the transition countries, need to play a major role in climate change mitigation. According to recent reviews, this message has come through as the abatement of greenhouse gas emissions is rising on the political agenda of many cities, particularly in the developed world [19, 20]. Greenhouse gas emissions can be reduced in many different ways by all sectors and at all levels of organization of human society. The overall urban form also plays a role. The “compact city” model has gained wide acceptance as a way to achieve more resource-efficient cities. It is based on the notion that urban density is related to energy consumption, as denser cities consume less energy per capita. This applies especially for car-based travel [21]. This relationship has been used as a powerful argument against urban sprawl, that is, the extension of urban areas through low-density developments. For instance, European cities have on average

expanded by 78% in area since the mid-1950s while their population increased by only 33% [22]. Large urban regions have developed, which spread far into the previous countryside and can be delimited by patterns of daily commuting [23]. Sprawl has even been observed in city regions with a decreasing population where people move out from declining inner cities to live in fringe locations [24].

While city governments rightly have given much and still-increasing attention to reducing greenhouse gas emissions, adaptation to climate change has been of much less concern until recently (see reviews in [19, 20, 25]). Moreover, activities are still mostly focusing on certain sectors such as increasing the capacity of the water supply and the sewer systems.

Following Crichton [26] (Fig. 1), the climate change-related risks of urban areas are a function of the (1) hazards, (2) exposure of the urban system to these hazards, and (3) its inherent vulnerability:

1. **Hazards:** By 2100, the average global surface temperatures are projected to rise by 1.1–6.4°C and the sea level is expected to rise by 0.18–0.59 m. Annual precipitation levels will increase in high latitudes and decrease in most subtropical land regions. Hot extremes, heat waves, and heavy rain events are



Green Infrastructure and Climate Change. Figure 1

Illustration of risk as a function of hazard, exposure, and vulnerability [26] (Adapted from [26] by [116] with permission from author)

expected to increase, while storms and cyclones will intensify [1]. Regions most vulnerable and/or exposed to climate change include the Arctic (high rate of warming), Africa (low adaptive capacity), small islands (exposure to sea level rise and storms), and the Asian and African “megadeltas” (dense populations combined with exposure to sea level rise, storm surges, and river flooding). Overall, three effects of climate change are of particular concern for urban areas:

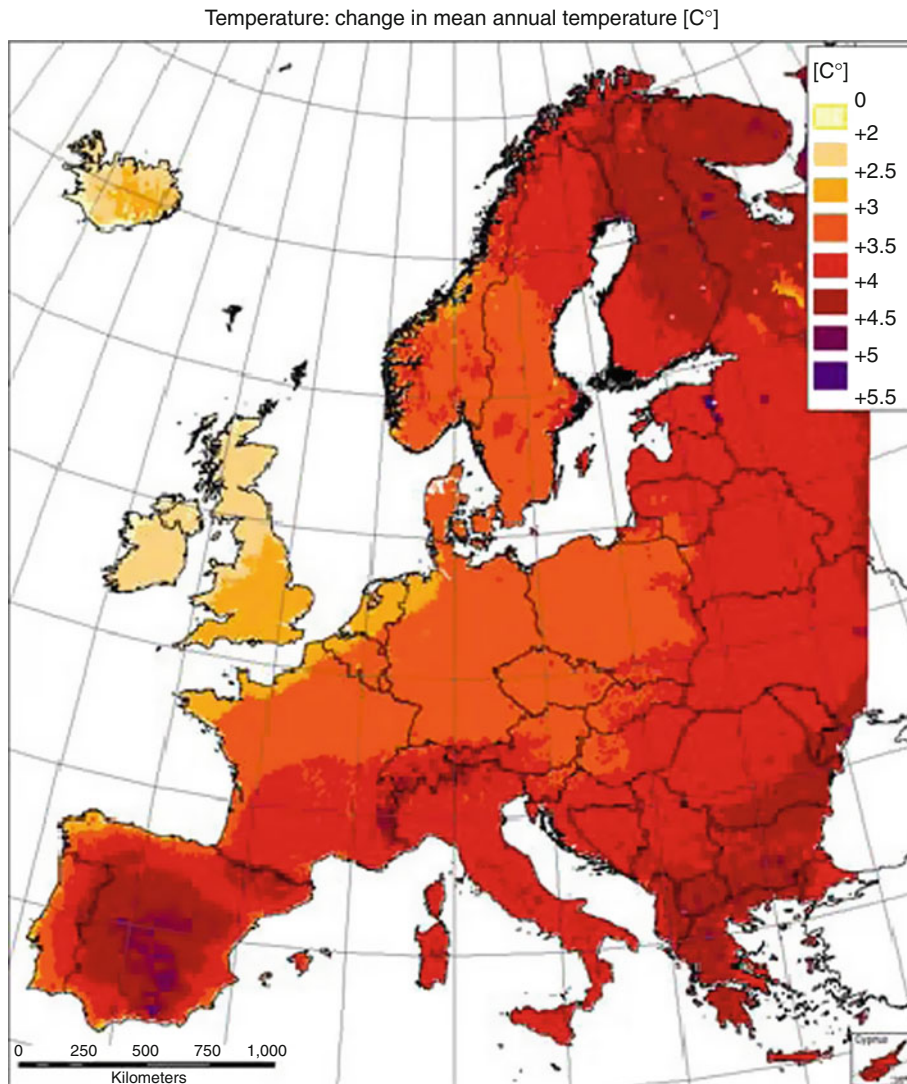
- Sea level rise and increase of storm surges caused by tropical storms
- Temperature rise, which intensifies the urban heat island
- Changing amounts and patterns of precipitation, which increase the risk of drought, on the one hand, and pluvial and sewer flooding and landslides from rainstorms, on the other hand

Taking Europe as an example (Fig. 2), temperatures will increase in particular in the North and the South but in more moderate ways in Central and North West Europe. In the North, temperatures and precipitation will particularly rise in wintertime, while the South will be affected by hotter and drier summers. Not all of these changes will be negative to living in the city. For instance, human winter mortality and heating demands are expected to decrease in northerly areas. Also, the summer half year will become more beneficial for outdoor activities in the north. Yet, even for countries such as Sweden, it has been estimated that these beneficial effects are outweighed by increased mortality due to heat waves and increased energy demands for cooling during summertime [27]. In particular, it needs to be highlighted that it is less the general change of temperature and precipitation patterns that are of concern but the increasing frequency and intensity of extreme events such as hot spells of weather and droughts, more frequent thunderstorms that bring large quantities of rainfall in a short time, and perhaps also a higher likelihood of heavy storms, although this is far less certain. For instance, it has been predicted that the number of so-called tropical nights will rise. Tropical nights are nights where the air temperature does not drop below 20°C. Under the IPCC scenario A2, it is predicted that the night temperatures in central Europe in 2080 will be similar to current levels in Spain [28].

2. *Exposure*: Urban areas are often located in zones particularly exposed to climate change hazards such as storm surges, river floods, and landslides. As an example, 13% of the world’s urban population live in the low-elevation coastal zone (<10 m above sea level) [29]. Almost two-thirds of urban settlements with a population greater than five million are located in this zone, including the densely populated deltas of Ganges-Brahmaputra (Kolkata, Dhaka), the Nile (Cairo), and the Yangtze River (Shanghai). These areas are at particular risk from flooding when high tides combine with storm surge and/or higher river flows. Yet, not all areas within coastal cities are equally at risk. In a developing world context, particularly the informal settlements of the urban poor are exposed to flooding.

3. *Vulnerability*: Urban vulnerability to climate change is multifaceted and related to the physical, social, economic, and environmental characteristics of urban areas. Various concepts of vulnerability have been proposed by the research community on climate change and disasters [30, 31]. The emphasis is thereby either placed on vulnerability, mainly as a result of physical factors, for example, location and layout of the city (“outcome vulnerability”) or of its social makeup, that is, the condition of people that enables a hazard to become a disaster (“context vulnerability”). Adoption of either one of the concepts has consequences for the measures to be taken to reduce urban vulnerability.

An obvious reason for the vulnerability of urban areas to climate change is the concentration of people, infrastructures, and economy. Further, the urbanization process per se increases the vulnerability. Urbanization fundamentally alters the earth’s surface by replacing vegetated or otherwise open land with built land, and urban areas may be characterized as open ecological systems with a high throughput of energy and matter. Consequently, ecosystem processes are strongly impacted, such as modification of local and regional climates, air quality, hydrology, soils, and flora and fauna [32]. Annual average air temperatures in big cities are already 1–3°C higher than in the surrounding countryside. Thus, already today, climatic conditions in urban areas may correspond to anticipated climate change. However, urban areas’ climate will be further modified by climate change and stormwater runoff will strongly increase [25].



Green Infrastructure and Climate Change. Figure 2

IPCC scenario A2 for Europe [117]. Absolute change in mean annual temperature between control periods 1961–1990 and 2071–2100, under the IPCC SRES scenario A2. The maps are based on PRUDENCE data (<http://prudence.dmi.dk>), and processed by the European Commission's Joint Research Centre, Institute for Prospective Technological Studies, within the PESETA study (<http://peseta.jrc.es>). ©European Union, <http://eur-lex.europa.eu/>

Planning: The critical role of land-use planning and management in climate change mitigation and adaptation has been recognized by the Intergovernmental Panel for Climate Change [1].

Urban planning and land management can mitigate the severity of hazards and reduce the levels of exposure and vulnerability [33]. Land-use planning can reduce transport demands and building regulations can

reduce the need for house heating or cooling [20]. Directing urban development away from floodplains can greatly reduce the risks of being affected by increasingly frequent and severe floods. As will be argued in the following sections, planning can increase the resilience or capacity of urban areas to cope with climate change by strategic planning of a green infrastructure. As Lindley et al. observe, “different scales of planning

from macro-scale land-use planning to micro scale urban design are both important to this process, responding to the different scales over which risk and vulnerability are expressed” [33, 34]. Integration of climate change-induced risks in urban planning is, however, only beginning to emerge [35, 36].

Urban areas are densely built, densely populated, have high land costs, and must support multiple functions. This calls for innovative solutions to the design of urban areas that are inclusive and synergetic. It is a challenge that can, at least partly, be addressed by transferring landscape design principles into the design of the city.

There is ample evidence of the beneficial role of urban green space in providing ecosystem services [37–39]. Green space can mitigate the urban heat-island effect and reduce stormwater runoff [40, 41]. Trees and shrubs also sequester carbon, and if the amount of living biomass in a city is permanently enlarged, the overall carbon balance of urban areas will be improved [42]; however, the contribution is likely to be only of marginal importance. The question that remains to be answered is to what extent cities can adapt to the expected large alterations of the environment driven by climate change through planning and design of an urban green infrastructure? Green infrastructure is an emerging concept increasingly gaining acceptance in the fields of urban planning and design as a way to promote multiple ecological and cultural processes in the city [9, 10]. Infrastructures facilitate multiple functions in the city including transportation, communication, energy, and water services. Most existing infrastructures are “grey infrastructures” such as roads and sewers. The “green infrastructure” complements these infrastructures by providing essential ecosystem services such as climate regulation, stormwater management, biodiversity, and social services such as recreation and aesthetics [9, 10]. Has green infrastructure, similar to other infrastructures such as transport and supply infrastructures, the ability to maintain the viability of the city as such by the provision of ecosystem services and increased resilience [10]?

The Role of Green Infrastructure for Adaptation of Cities to Climate Change

Adaptation of cities to climate change refers to the implementation of means to reduce the impact of

climate change. By timely preparing the city to face the impact of a changing climate, economic and social costs and potential damage costs can be reduced while synergies can be realized, for instance by developing more attractive and livable cities [43]. Development of the green infrastructure is an effective strategy both for adaptation to and mitigation of climate change.

Adaptation to Rising Temperatures – The Heat-Island Effect

The local climate in urban areas differs from the surrounding open land [44]. Of particular relevance in the context of climate change is the increase of air temperatures in urban areas by 1–3°C on annual average compared to the surrounding open land. On warm summer days, the difference can be up to 5–12°C [45]. This is commonly referred to as the urban heat-island effect. The intensity of the heat-island effect is dependent on the size of the city: the larger the city the stronger the heat-island effect. This is relevant as many urban areas in the world are strongly growing, and the resulting increase of air temperatures in inner cities may be stronger than that caused by global climate change. However, climate change is expected to further intensify the heat-island effect. For instance, a modeling study predicted intensification of the heat-island effect for London by 0.5°C by the 2050s [25].

In warmer cities, it will be more difficult to work and sleep properly without cooling and ventilation systems. This will increase energy demands; it will cost money and cooling systems are only available to people who can afford it. In the USA, peak urban electric demands were estimated to rise by 2–4% for each 1°C rise in daily maximum temperature. The additional use of air-conditioning is responsible for 5–10% of urban peak electric demand in the USA, at a direct cost of several billion dollars annually [46].

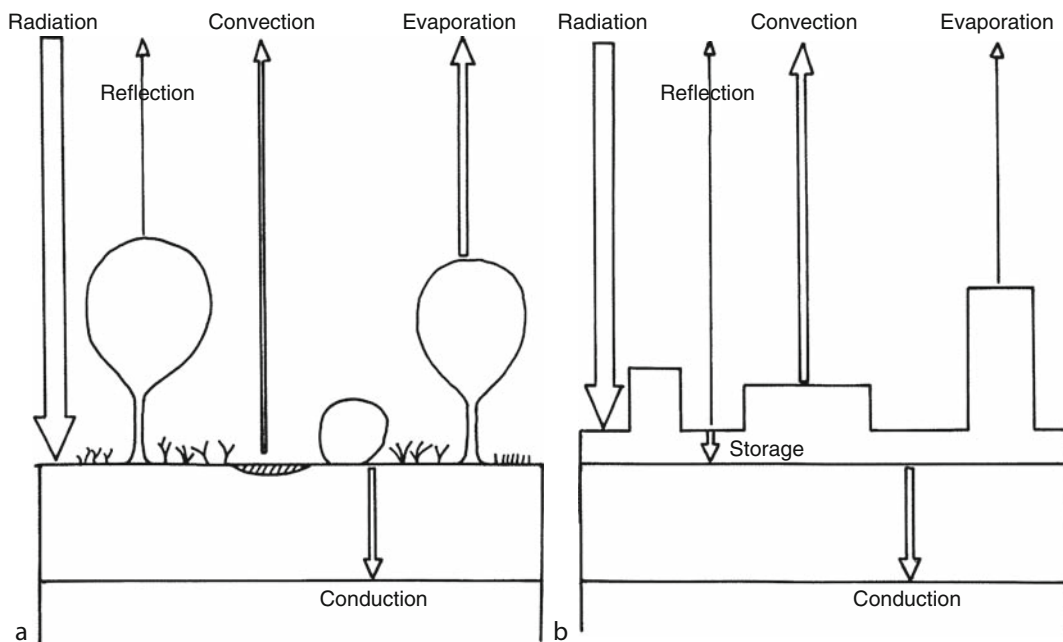
Increased temperatures in urban areas can have negative consequences for human health and well-being. In particular, small children and elderly people as well as persons suffering from cardiovascular diseases are at risk from hot spells of weather. An example is the extreme summer of 2003 in Europe where more than 70,000 deaths in excess have been reported [47]. Most casualties occurred in urban areas where most people live and the heat-island effect intensifies.

As outlined above, such hot summers with extreme heat waves will become more frequent and intense in the future [28]. In addition to direct impacts through raised temperatures, the heat-island effect also deteriorates air quality. For instance, the heat-island effect in Los Angeles, California, has been estimated to increase ozone concentrations by 10–15 [48]. It should be noted, however, that climate change impacts on urban areas will play out very differently according to location and type of city and a decline of winter mortality can be expected in urban areas in colder climates.

The heat-island effect is a result of the altered climatic balance in urban areas and it is caused by a number of factors, such as the reflection, storage, and convection of solar energy and emission of heat from anthropogenic process (Fig. 3, [41, 45, 49, 50]). The main reason, however, is the replacement of vegetated, evaporating surfaces by built and paved surfaces. Vegetation consumes much of the solar energy for evaporating water. This imbalance is fortified by the fact that buildings, roads, and pavements are made of materials with a higher capacity to accumulate and

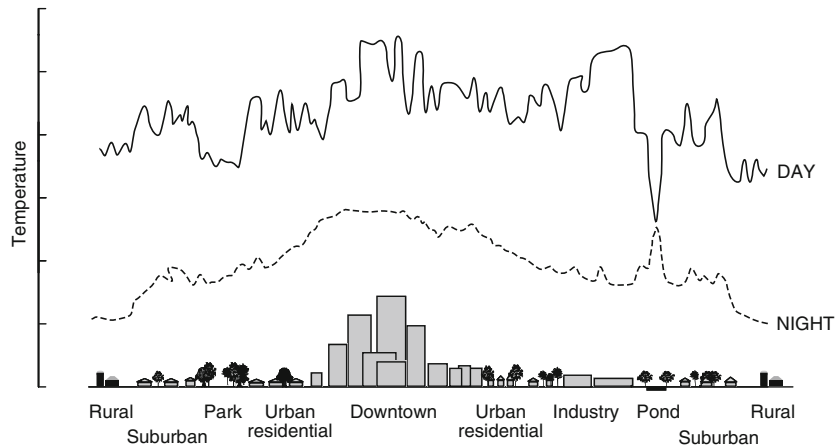
store heat than living tissues. Measurements show surface temperatures do not decrease continuously along urban to rural transects but vary strongly within the urban area depending on the physical characteristics of the different land uses along this gradient (Fig. 4). Temperature peaks can be observed in densely built parts of the city, while temperatures are considerably lower in well-greened areas. Therefore, there is not one heat island but in reality a city consists of a small-scale archipelago of heat and cool islands which corresponds well to patterns of urban form.

Although air and surface temperatures can be lower in green spaces than in the surrounding built areas, the effect varies depending not only on the location, size, and design of the green space but also their management (e.g., whether lawns are irrigated). In a large park, annual average air temperatures may be 1–2°C lower than the surroundings. During clear windless nights, temperature differences can be as big as 5–6°C. Moreover, heat is trapped in narrow street canyons of densely built areas. However, parks must have a certain size to develop a distinct climate. According to measurements



Green Infrastructure and Climate Change. Figure 3

Energy transfers contributing to the urban heat-island effect: (a) the situation in the open land, (b) the situation in densely built-up areas [40] (Reprinted from [40]. Copyright (2001), with permission from Elsevier)



Green Infrastructure and Climate Change. Figure 4

Variations of surface temperatures in urban transect. (Modified from [118] with permission from Actionbioscience.org the author)

in public green spaces in Berlin, temperature differences could only be found when the park was bigger than approximately 3.5 ha [51].

Larger parks can also moderate air temperatures in adjacent built areas. The study on green spaces in Berlin (Germany) established that very large parks such as Tiergarten (212 ha) can lower air temperatures on clear summer days with low wind speeds (<2 m/s) up to 1,300 in lee and up to 200 m windward [51]. Yet, while these figures may be impressive, the effect is much less for smaller green spaces where effects may be limited to distances of 100–200 m. The cooling effect also depends on the surrounding built structures. Corridors in the main wind direction may lead cool air from the park into adjacent built areas whereas closed built structures around the park block the cool air [52]. Therefore, mitigation of the urban heat island cannot rely on single large green spaces. Instead, a dense network of green spaces should permeate the built fabric. Moreover, greening of the urban matrix is crucial to provide for comfortable climatic conditions close to where people live and work. Air temperatures clearly differ between urban land uses dependent on the degree of greening. Even in very densely built areas, shade trees can be planted in streets to control climates at site levels while roof and wall greening reduce the heating up of built surfaces and heat storage. Trees are particularly effective in cooling the city as they have a shading effect.

Therefore, parks with a high provision of trees and urban forests are usually the coolest outdoor areas during daytime.

While there is ample evidence from urban climate studies to establish the relationships between urban form, green space, and climates at various scales – from city to streets – research on the likely consequences of climate change on urban climates and the potential of green infrastructure to mitigate these effects are still scarce. One of the few exceptions is a study undertaken in Greater Manchester, UK. In the project “Adaptation Strategies for Climate Change in the Urban Environment” (ASCCUE) [53, 54], a modeling approach was employed to explore the impacts of climate change on urban temperature patterns and establish the relationships with urban form. In Greater Manchester, a large conurbation of approximately 2.5 million inhabitants on a surface area of approximately 1,300 km² was selected as it suitably represents the different forms of built and open spaces that are typical for large urban areas in the UK. Greater Manchester is also an interesting case as it undergoes a rapid process of transformation from a former industrial to a postindustrial city where land-use dynamics are high. It also offers opportunities for development of a green infrastructure for climate change mitigation and adaptation.

In the ASCCUE project, the researchers developed an approach for the assessment of climate-related risks

with reference to three exposure units: built, green space, and human comfort. However, only the results from research on urban green space will be reported. This study consisted of three main steps:

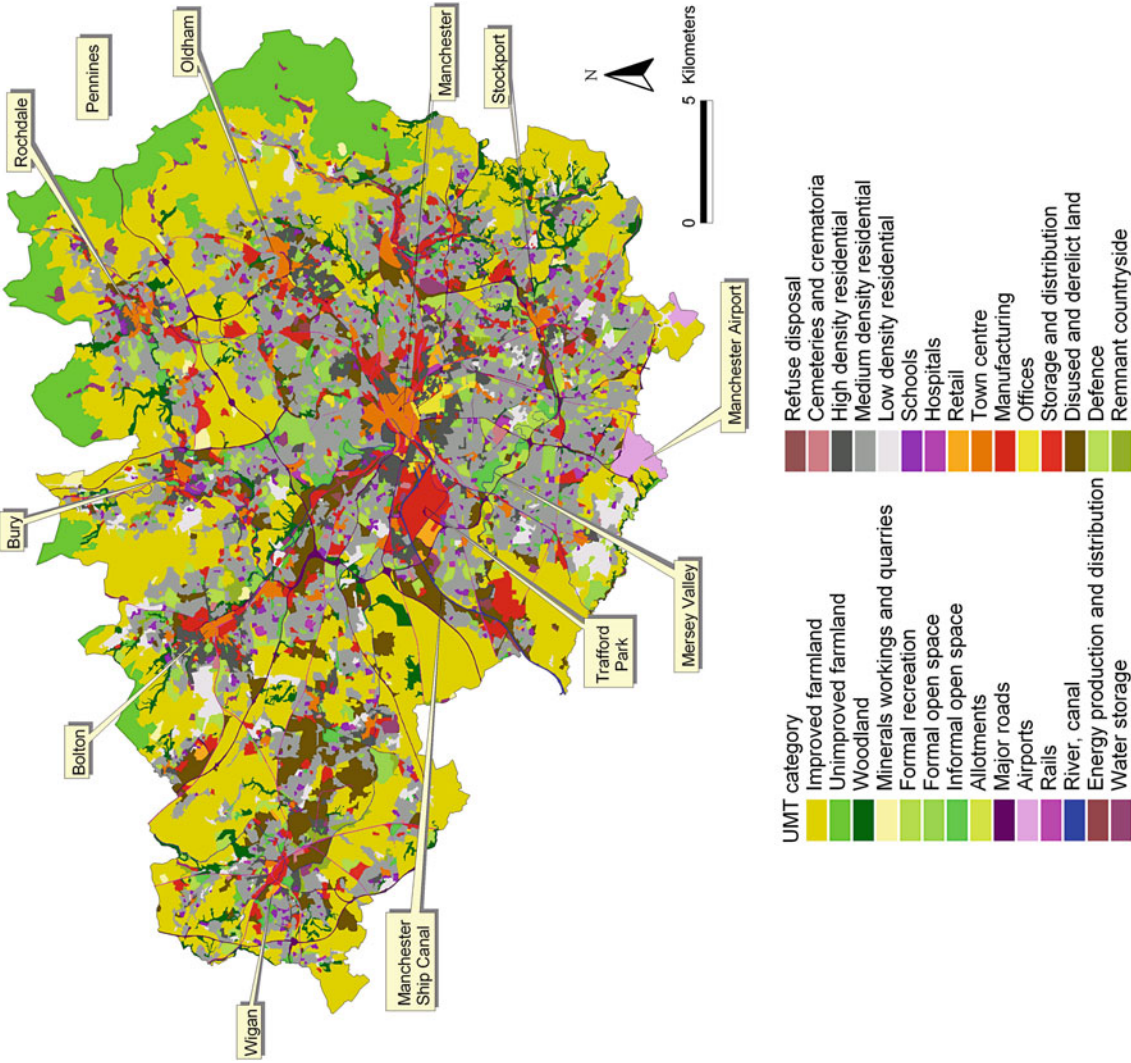
1. Characterization of the urban environment based on mapping of urban morphology types from aerial photographs [55]: Overall, the area was classified into 29 “urban morphology types” (Fig. 5). These were further characterized by nine land cover attributes, including cover of built, paved, and vegetated areas by interpretation of aerial photographs. For the latter, detailed estimates were derived for trees, shrubs, rough grasslands, lawns, and arable land.
2. Modeling surface temperatures (and stormwater runoff, which will not be reported here) based on an energy exchange model [40, 54]: Surface temperatures are a useful climate indicator for urban planning as information can be derived from remote sensed imagery or – as was the case here – from models with complete spatial, highly resolved coverage of urban areas whereas measurement of air temperatures is confined to point data or transects from where it is difficult to spatially extrapolate. Moreover, surface temperatures can be taken as a proxy for mean radiant temperatures, which have a strong influence on human comfort and health in outdoor environments, especially on hot days with little wind [56]. For instance, impacts of the heat waves in 2003 in Paris showed that increased levels of human mortality were closely related to heat islands of surface temperature derived from remote sensed data [57]. In the Manchester study, the surface temperature model’s main input were land cover data, on the one hand, and climate data from the local weather station, on the other. Climate data were obtained from a parallel project, where climate scenarios were derived for the weather station of Manchester airport for 2020, 2050, and the 2080s [58, 59]. Only results for the most extreme emissions scenario called “2080 high” will be reported here.
3. Urban development scenarios to compare the impacts of “green” versus “densification” strategies under climate change until 2080: In the green strategy, cover of vegetated and water surfaces would be increased by 10% in terms of surface cover while the

densification strategy would lead to a reduction in cover of evapotranspiring surfaces by 10%.

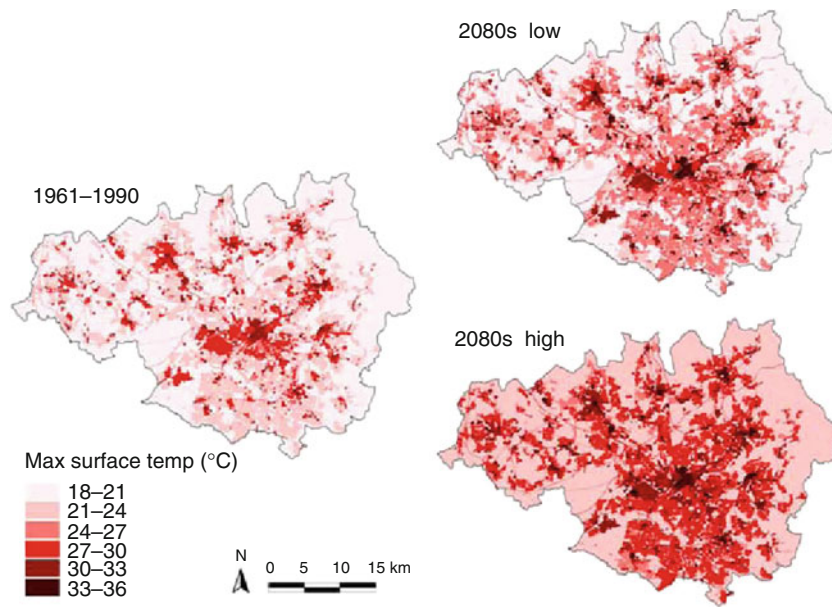
The share of green space varies between urban morphology types. In the inner urban areas, the share of green space is usually low (less than 20%); in single-family housing estates, the provision of green space is 50–60%. Overall, formal and informal green spaces (including woodlands) cover approximately 10% of the city surface; another 10% is covered by vegetated areas on derelict land of former industrial sites. Yet the largest green space resources are residential areas where approximately 40% of all vegetated areas can be found. Large differences, however, could be observed between densely built-up residential areas (e.g., terraced houses), and the medium- and low-density types. These differences generally coincide well with the socioeconomic status of the areas with deprived areas being at the lower end of green space provision. Moreover, they often lack in mature trees with large crowns, which are particularly effective in reducing surface and air temperatures via their shading effect.

Surface temperatures greatly differ between the urban morphology types on hot summer days, and these differences are clearly related to the cover of green space (Fig. 6). The highest surface temperatures were identified in the inner city, where buildings, roads, and pavements predominate and where the provision of parks, trees, and other types of vegetation is lowest. Temperatures were also very high in densely built residential areas and industrial and commercial estates. In well-greened neighborhoods with single-family houses, the temperature is considerably lower, whereas woodlands offered the coolest places. The study measured 32.1°C in the urban center, while the temperature in a large park was only 18.4°C.

In the Manchester study, climate scenarios for the year 2080 predicted a rise of air temperatures of more than 4°C during heat waves. The rise in temperature is significantly lower in areas with a high provision of green spaces. While surface temperatures would increase by 4.3°C in the city centers, they would only rise by 3.1°C in low-density residential areas. This shows that the cover of vegetated areas has potential to buffer the effects of climate change to some degree. These results compare favorably with the predicted intensification of the heat island in London mentioned before [25].



Green Infrastructure and Climate Change. Figure 5
Map of Greater Manchester divided into 29 urban morphology types (Reprinted from [55]. Copyright 2008, with permission from Elsevier)



Green Infrastructure and Climate Change. Figure 6

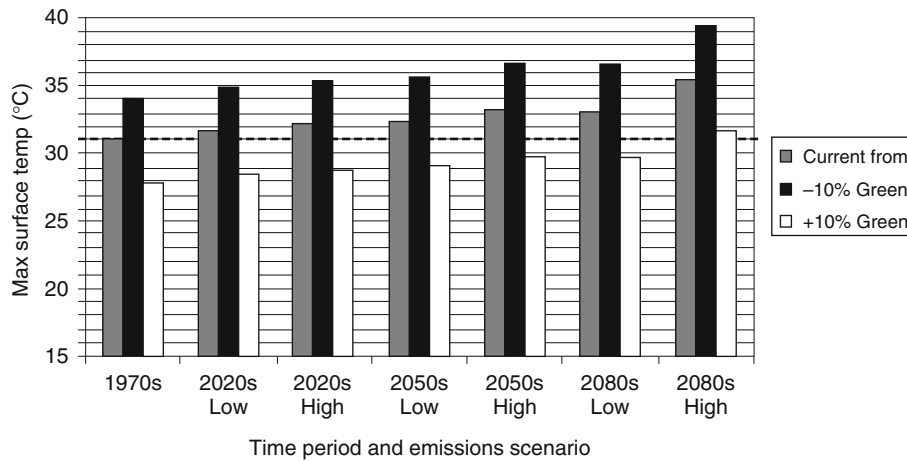
Maximum surface temperatures in Greater Manchester on a hot summer day in 1961–1990 and the 2080s low- and high-emission scenarios [54] (Reprinted from [54]. Copyright (2007), with permission from Alexandrine Press)

The climate scenarios for 2080: the “green strategy,” where the provision of green space in the densely built town centers was increased by 10%, and the “grey strategy,” where the provision of green space was reduced by 10%, resulted in further support for the important role of green space to increase the coping capacity of urban areas to climate change. It was estimated that 10% more green space would almost compensate for the rising temperatures expected to result from global warming by 2080 even in the worst scenario. In contrast, by reducing the provision of vegetated areas by 10%, the temperature in dense urban areas would rise by 8.2°C (Fig. 7).

It will be very challenging, though, to introduce more green space into densely built urban centers. However, the dynamics of urban areas should not be underestimated: current uses become obsolete, buildings are demolished, and entire neighborhoods are renewed. This may give opportunities for development of a green infrastructure if strategies are clearly defined and supportive to the wider agenda of urban renewal. In particular, this may be feasible in cities with a shrinking population [24] where creation of green spaces can also increase quality of life and thus help to

reverse the tendency of decline. In the city of Leipzig, the potential of developing urban woodlands on derelict inner urban sites is currently explored [60]. The project revealed that overall there are more than 500 sites within the built area covering over 1,100 ha. Many sites are located in deprived areas where there is a lack of green space and environmental quality is poor. Therefore, these derelict areas offer a great potential to improve quality of life in these areas. Moreover, they would allow developing a network of strategically placed “cool islands” in otherwise still densely built areas for climate change adaptation.

Where pressure on the land is high, and urban densification rather than urban sprawl is sought, it may be more difficult to develop a green infrastructure and other strategies need to be adopted. Even there, opportunities may exist. For instance, in the city centers of Manchester, about 37% of the city is covered by buildings but 40% by paved spaces of roads, car parks, etc. Thus, it is not the lack of space as such but the intensive use and in particular the occupation by car-based transport which is problematic. Evidently, it is difficult to change this situation but adaptation to climate change will require more radical measures to



Green Infrastructure and Climate Change. Figure 7

Maximum surface temperature for a hot summer day in inner city areas of Manchester, with current form and when 10% green cover is added or removed [54] (Reprinted from [54]. Copyright (2007), with permission from Alexandrine Press)

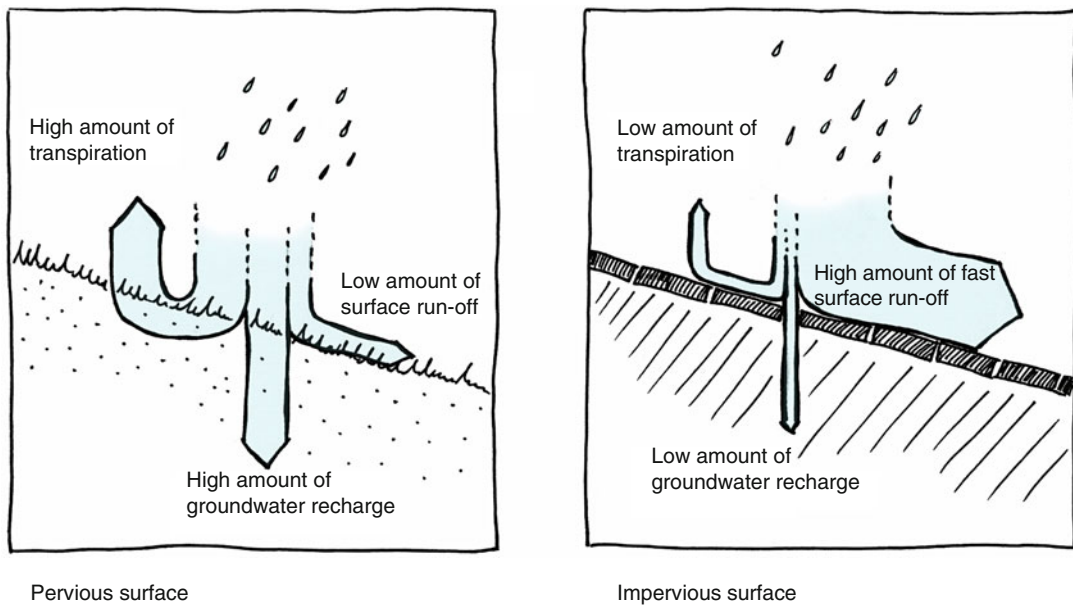
be taken for the reorganization of transport in city centers in order to create the space necessary for green space. Here, synergies with climate mitigation are obvious. In addition, green roofs and walls or implementing vegetated permeable pavements should be promoted.

Stormwater Management – From Sewer-Based to Landscape-Based Systems

The ongoing sealing of surfaces due to buildings, roads, and other infrastructure in urban areas leads to a change in the water cycle (Fig. 8). Instead of naturally high transpiration rates and water infiltration for groundwater recharge, large amounts of the precipitation become surface runoff. In natural environments, about 10% of precipitation will run off on the surface and 50% infiltrate, whereas figures are reversed in dense urban settings with about 50% surface runoff and 10% infiltration – of course very much dependent on the specific degree of surface sealing [4]. This leads to a higher and more rapid peak discharge in urban areas, which needs to be handled by the existing sewer systems.

Due to fast urban development and the aging of the systems, already today many existing sewer systems have insufficient capacity. Increase of water-impervious surfaces has been dramatic in urban areas

of the Western world. For instance, a study of 26 cities across Europe revealed that the size of urban areas increased by 78% on average between 1950 and 1990 while population increased only by 33% [22]. This low-density development has greatly enlarged the area of water-impervious surfaces for buildings and roads. In addition, densification of existing built areas increases the amount of water-impervious surfaces. Such densification may take different forms, from building over derelict or other non-built areas, infill development, for example, by building houses on vacant land or subdividing large gardens into several smaller parcels. Paving of front gardens to create parking spaces can lead to significant increase of water-impervious surfaces. Research in 11 residential areas in the Liverpool conurbation revealed that approximately 5% of the surface area was converted from vegetated to paved, between 1975 and 2000. The runoff from a 10 mm rainstorm event was modeled to increase on average by 4% [61]. While these changes may appear to be small, they can lead to significant problems for the sewage system when accumulated over an entire sewer catchment. In another UK-based study, increase of water-impervious surfaces due to conversion of front gardens was even bigger: A 13% increase in water-impervious surfaces was detected in a 1.16 km² suburban area of Leeds in northern England over the 33-year study period between 1971 and 2004 [62].



Green Infrastructure and Climate Change. Figure 8

The water cycle in natural and sealed areas. (Adapted after [119] with permission from the author)

An increase of stormwater runoff by 12% was modeled as a consequence.

The increase of stormwater runoff leads more and more often to sewer overflows. This can either be uncontrolled, resulting in the flooding of urban areas, or the overflow happens controlled into nearby streams and other receiving water bodies. In this case, the ecological consequences can be dramatic.

Climate change will strongly exacerbate these problems. In the Manchester study, it was predicted that the amount of precipitation from rainstorm events will increase by 56% until 2080 in the high-emissions scenario. This would lead to 82% more runoff because of limited water retention and infiltration capacity of the soils [54]. Protection of green space on soils with a high infiltration capacity such as sandy soils will be a major task for urban planning to avoid a further dramatic deterioration of this situation. An increase of green space cover by 10% – as explored in the case of heat-island mitigation above – would reduce surface runoff during a rainstorm event by approximately 4%, and thus not suffice to solve the problem. However, implementing on-site retention and infiltration techniques can be a way out, as will be explored in the following section.

The northern European country Denmark will be taken as an example. Comparing regional precipitation data from 1979 to 1997 with 1997–2005 suggests a 10% increase in rain intensity for a 10-year design storm [63]. Model simulations project that extreme rain events will further increase by 20% or more during the next 100 years [64]. To comply with more intensive rainstorms, the drainage capacity needs to be increased by approximately 30% in a 100-year time perspective [65]. Including uncertainties related to urban drainage modeling and urban densification processes, the expansion of drainage capacity is estimated to be 78% higher than present level [65]. This is only to maintain the same flood risk as today, that is, surcharge accepted every tenth year in combined sewers collecting stormwater runoff and wastewater from households and industries in the same pipe system. For the country as a whole, the cost of increasing drainage capacity by conventional means (i.e., installing bigger and more sewers, basins, and pumps) is estimated to be \$2.7 billion USD in a 30-year time perspective [66]. Combined with sewer rehabilitation works resulting from aging infrastructures, the total cost accounts to \$14.6 billion USD [66]. With a total urban population of 4.5 million people, this corresponds to investment needs in

the range of \$3200USD per urban capita. Still, there is no certainty that the predicted level of climate changes are reliable and that the one-off investments in new pipes (relying on long depreciation periods) will provide sufficient drainage capacity in the future. If no action is taken, the risk of floods will increase, which will impact people's assets and livelihoods.

Large challenges lie ahead in adapting cities to more extreme rain events. Conventional solutions are costly and possibly unsustainable long term [67, 68]. Another solution is to imitate the natural water cycle with delay, infiltration, or evaporation of the stormwater as close to where it falls as possible. Single measures for such on-site stormwater management are elements like green roofs, permeable pavement, swales, ponds, and infiltration trenches. This strategy is also referred to as landscape-based stormwater management, Sustainable Drainage Systems or Sustainable Urban Drainage Systems (SUDS), Best Management Practices for Stormwater (BMPs), Low Impact Development (LID), or Water Sensitive Urban Design (WSUD) [69, 70].

Sustainable urban drainage systems (SUDS) can be used as a full stand-alone solution typically employed in new urban developments or retrofitted into the existing city to provide more drainage capacity in the area, in which case it is more likely to be complementary to the existing sewer system. SUDS further provide the opportunity for comprehensive solutions to suspend recreational and spatial utilization of the water as well as increased biological diversity and well-functioning urban drainage systems that contribute to increased groundwater recharge. This type of solution is expected to include economical, environmental, and recreational advantages.

SUDS have been implemented on various places in the world and can include a wide range of different measures. Concepts of on-site stormwater management have growing acceptance in Australia [71], Germany [72], the Netherlands [73], and the UK [74] though still not widely and commonly applied in all settings.

A range of studies show the positive impact that a new stormwater management approach can have on cities. One such example can be found in Malmö in Sweden [75], where a combined green roof and pond system is retrofitted into the Augustenborg neighborhood. Green roofs are one measure out of a wide range

of possibilities for stormwater management. For instance, Bengtson [76] established that extensive green sedum roofs with a 3 cm thick substrate layer are expected to retain 10 mm of rain before discharging water. Models run in parts of the Augustenborg neighborhood indicated that a 31% extensive green roof cover in a certain area can reduce the peak flow by 64% for a rain event statistically happening twice a year, and 27% for a rain event reoccurring once in 5 years. In addition, the runoff volume is reduced by 52% and 18%, respectively, for the same return periods [77].

The implementation of SUDS is expected to improve the quality of life in the cities and has furthermore a range of technical as well as ecological advantages. It leads to an improved performance of the wastewater treatment plants, as the wastewater is less diluted, to a reduction of energy demand for wastewater treatment and furthermore to a decrease of combined sewer overflows. Stormwater can be harvested for supply purposes, such as laundry and toilet flushing, and thus reduce the strain on other freshwater sources.

Moreover, infiltration of stormwater can contribute to recharge and maintain groundwater sources in urban areas. Infiltration of stormwater in soil layers that are hydraulically linked to local streams can support the stream's base flow. A lower peak flow reduces the risk of flooding, erosion, and uncontrolled pollution of receiving water bodies. In cities, the groundwater level is commonly lowered as a result of extraction of groundwater for supply purposes, drainage of natural wetlands, and little stormwater infiltration due to few permeable surfaces. A general decrease in the groundwater in an area can result in a lower water table in surrounding wetlands and negatively affect the flora and fauna. It should be noted that climate change is predicted to increase the overall amount of precipitation, which would in general increase groundwater levels in Northern European countries such as Denmark.

Finally, SUDS support the use of rainwater as an asset benefiting city life. People are attracted to water whether it is children playing in a puddle or adults who are happy to pay more for a house with a view to waterfronts than other places. Water increases the attraction of the environment and can contribute to a greener and richer plant and animal life.

Apart from the advantages, the implementation of SUDS bears also some risks, such as groundwater pollution from polluted soils or infiltration of runoff from heavily trafficked areas, or the unwanted flooding of low-lying areas or basements. These risks need to be known by planners and engineers and proper avoidance strategies and monitoring needs to be implemented together with the SUDS.

SUDS are a relatively new concept and not as many experiences exist as goes for conventional sewers. Some of the major aspects that emphasize the difference between conventional systems and on-site stormwater management systems are, for example, the dimensioning, the work across scales, the context dependence and the need for new collaborations, and a new way of thinking.

Regarding urban drainage systems, a good solution must cover and respond to different conditions. It must be suitable for the everyday functionality of the city and appropriately manage the risk of flooding, for example, by redirecting water to areas where the damages and negative impacts are smallest.

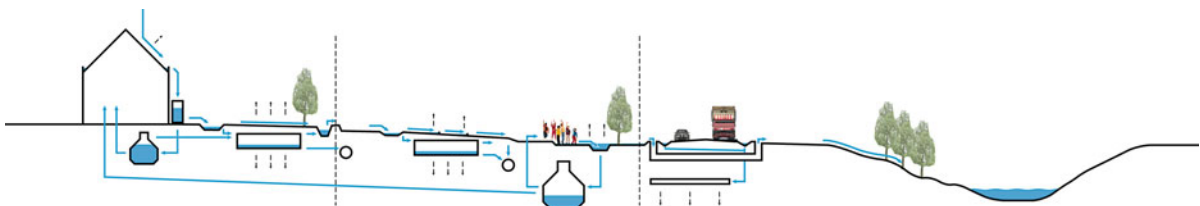
Furthermore, stormwater management planning is closely linked to environmental management for entire watersheds. Up to now, most projects have been implemented only at site level with little connection to the surrounding landscape [78]. Little knowledge has been developed regarding large-scale landscape-based stormwater management systems and their potential positive and negative impacts on other aspects of urban land use and possible integration with other urban development goals. An interdisciplinary case study with the goal to develop a suitable strategy for large-scale SUDS retrofit in the eastern parts of Copenhagen (Denmark) [79] suggested that there is a need to place small-scale projects for on-site

stormwater management within the context of larger-scale planning for entire watersheds. As an aim, a system of point measurements should be connected by linear elements and zones, and by this create a new green and blue water infrastructure network. Such a system design of connected disconnections would support the capacity of the single sites across special and temporal scales. Single site measures would treat the stormwater independently on a daily basis, while in case of an extreme storm event, streets and green corridors function as emergency overflow systems. Figure 9 is developed as part of the study in Copenhagen and illustrates an interconnected system of SUDS elements covering site-level, district-level, and city-level measures. It points toward the potential of urban green infrastructure to support the planning and design of large-scale landscape-based stormwater management systems, as integrated sites for retention, infiltration and evapotranspiration, and/or conveyance of excess runoff that can be linked to the planning of gardens, parks, and green corridors.

Such a citywide water system requires close collaboration between the different stakeholders. Green space management and sewer management must merge to ensure successful SUDS implementation, and as water does not know administrative borders, a planning system that brings all stakeholders to the table is required.

Flood Management

Sea level rise will challenge many coastal urban areas. In addition, increased runoff will raise river water levels and increase flood risks. More impervious urban areas and more intensive rain events will exacerbate the situation. The conventional way is to manage rivers by means of channelization, raising the dikes and



Green Infrastructure and Climate Change. Figure 9

Sketch of an interconnected system of SUDS elements (illustration: O. Fryd)

installing dams. A landscape-based approach to coastal and river flood management is an emerging concept currently being explored in countries like China [80], Germany [81], India [82], the Netherlands [83], Singapore [84], and the USA [85].

In quite many cities, projects of river restoration have been implemented in the last decades, for example, revitalization of the Los Angeles River (USA) [86], Cheonggyecheon Linear Park, Seoul (South Korea) [87], Akerselva, Oslo (Norway) [88], the Aarhus River (Denmark) [89], and the river Isar in Munich (Germany) [90]. Climate change may not yet have been a strong political driver of these projects, but the outcomes are certainly highly relevant for adaptation of cities to climate change via the green infrastructure. As an example, the restoration of the river Isar within the built area of Munich will be briefly presented here. Munich is a city of some 1.3 million inhabitants located on the banks of the river Isar. The river originates in the Alps, and the city's history has been closely connected to the river as a source of water, energy, and a transport route, and the Isar is just as iconic to the city as the Alps nearby or the Oktoberfest. While the medieval city kept a respectful distance to the wild river, the Isar was channelized from the nineteenth century onward for

flood prevention and energy use. The floodplains were partly built over. Still, the river is the backbone of the city's green structure with famous parks in the former floodplain.

In the 1990s, a major flood almost caused a catastrophe with river boards being filled to the brink. This, together with new regulations whereby safety thresholds for flooding were increased, caused a radical rethinking of the approach to flood management. It was realized that flood problems could not be solved alone by further increasing the height of the river dams. Moreover, this would have caused major negative impacts on green spaces along the river which are highly valued for recreation and therefore, it would have been difficult to get political and public support for a hard engineering solution. Instead, a strategy was adopted whereby the river bed was broadened where this was still possible to provide more space for the water. The riverbanks were remodeled to enhance access to the river for recreational purposes (water quality has been improved due to new sewage treatment systems of municipalities upstream) and to promote habitats for wildlife (Fig. 10). The river Isar restoration can therefore be considered as an example where a politically contentious situation was turned



Green Infrastructure and Climate Change. Figure 10

River Isar in Munich during a flood after restoration (photo: S. Pauleit)

into a win-win solution by adopting a multifunctional strategy. It should be noted that this was possible due a number of preconditions that had paved the way. Among these was a long history of steadily aggravating problems with the river ecosystem that necessitated taking measures along the entire river, the outstanding role of the river for city image, a society campaigning for the protection of the river that existed for more than 100 years, possibility to renegotiate the use of the river water for power generation due to expiry of contracts, and construction of sewage plants in municipalities upstream, which improved water to bathing quality. Opening of this window of opportunities was essential for realization of the project.

The examples of river Isar and others show that landscape-based flood management systems along coasts and rivers can be fully or partly developed as part of the urban green infrastructure. Landscape-based flood management systems reflect some of the same principles as SUDS by using hydrology as generator of sustainable urban form by specifically addressing flood risks, seasonal flow patterns, and tidal water dynamics as defining factors for integrated design. Much focus seems to be on the edges between cities and their water bodies, that is, a transition from hard edges to softer and more dynamic edges. This is characterized by a change from channelization and embankment to a focus (at least partly) on temporarily wet and dry transition zones. From an ecological perspective, such zones are expected to perform well as marshes, wetlands, mangroves, and riverbanks provide some of the most diverse natural ecosystems. However, they must be designed to also comply with urban needs such as public health and accessibility. Therefore, a multifunctional strategy is prerequisite to successfully developing the green infrastructure.

The Potential of Green Infrastructure to Mitigate Climate Change

Climate change mitigation relates to the reduction of greenhouse gas emissions. For this purpose cities have a series of potentials, for example, by reducing energy consumption rates, exchanging non-renewable by renewable energy sources, and densifying cities so the need for transportation is reduced. As an example, Melbourne, Australia, aims to be CO₂ neutral by 2020

[91]. Many other cities throughout the world are currently launching initiatives to reduce CO₂ emissions and to promote a more climate-friendly development.

To some extent, green infrastructure can mitigate greenhouse gas emissions, either by directly capturing and storing CO₂ or by reducing energy consumption of nearby buildings. It should be stressed, however, that the literature is sparse and that the mitigation potential of urban green infrastructure is not well documented. The chapter heavily, but not solely draws on Nowak [42], where the potential of urban trees to modify the urban environment has been reviewed.

CO₂ Sequestration and Storage

Trees and other types of vegetation take up CO₂ when they grow. A large, old tree can store about 3 t of carbon in the stem, branches, and roots [42]. This corresponds roughly to the amount of CO₂ emitted from driving 18,000 km in a medium-sized car (based on the assumption that a car emits on average 164 gCO₂/km) [92]. As such, the uptake in urban trees is very modest compared to the total CO₂ emission deriving from a city. For Chicago, it was estimated that the yearly sequestration of CO₂ in all the urban trees totaled approximately 140,000 t, corresponding to the CO₂ emissions of all car-based traffic from 1 week [42]. Further, the improvement is only permanent if the trees, when cut or dead, are replaced by new plantings and the dead biomass is used to replace fossil fuel. Simple decay will release the entire amount of stored CO₂ back into the atmosphere. In the balance of biomass CO₂ sequestration, the emissions of CO₂ from decay processes should also be subtracted.

Increasing the amount of green in a city will only marginally meet the mitigation challenge. The main task is still to reduce greenhouse gas emissions. Still, it can be reasonable to include urban trees in a climate plan and as part of the citywide CO₂ strategy as it might serve as part of a city-branding strategy. New York City has launched the “Million Trees Programme” to increase the amount of trees in the city by 20% as a means of carbon sequestration. At present, the urban trees in New York take up 42,000 t of carbon every year and stores a total of 1.35 million t of carbon [93].

Large tree size, longevity, and high growth rates are factors that have a positive influence on trees capacity

to sequester and store carbon. However, trees can only fulfill these functions when they are healthy and growing. In particular, street trees suffer from many stresses, and the average age of such trees has been estimated to be as short as 10–15 years in some cases [94, 95]. Moreover, recently planted trees often die due to bad site conditions and lack of or improper care [96]. Exact figures are rare but for instance, for Liverpool (UK) it was estimated that 39% of all newly planted street trees died within 5 years after establishment [97]. Yet, as a modeling study highlighted, street trees need to grow for at least 5–10 years before trees start to have a positive carbon balance because of the amount of carbon spent on their raising and management [98].

Urban trees are not the only way of greening the city. Lawns, shrubs, green roofs, and grass swales comprise some of the other options. Still, CO₂ storage capacity varies a lot and is mostly dependent on biomass stored in trees. A study of four residential neighborhoods in Liverpool showed a variation in carbon storage from less than 1 t per hectare to 17 t per hectare, with corresponding annual carbon sequestration rates from close to zero to 130 kgCO₂/ha [40].

CO₂ sequestration and storage in vegetation has a potential as an “image” of sustainability, but the impact of mitigating climate change is very modest in actual figures. To obtain larger reductions in CO₂ emission rates, urban greening can be used as a strategic tool to reduce energy demands, as presented below.

Reduction of Energy Demand for House Heating and Cooling

Hot weather in the summer period in presently temperate areas will increase the demand for air-conditioning in office buildings as well as in private houses. Yet, air-conditioning is energy demanding and expensive. Green infrastructure can potentially serve as a passive system, which can reduce energy demands for house heating and cooling.

Shading walls in summer time reduces the thermal load on buildings and hence the need for air-conditioning in warm climates. Trees should be best planted on the side where the afternoon sun hits the buildings (the southwestern side on the northern hemisphere). Deciduous trees are better than evergreen trees as the latter block the sun in winter when it warms up

the building. Placing evergreen trees at the side of prevailing cold winds during wintertime can also shelter buildings and thus reduce energy loss [42]. These are significant factors in particular when insulation of buildings is not at high standards. For the homes in the US it was estimated that energy use in house can be 20–25% lower with an optimum planting of trees around them [99]. Yet, when badly placed, trees may even increase domestic energy consumption.

Other model calculations from the USA showed that 10% more tree cover can reduce energy consumption rates for cooling by 24% in Sacramento and 12% in Phoenix [100]. Most of the cooling energy savings were attributed to the effect of evapotranspiration and only 10% to the direct effect of shading and wind shielding. In the more northern city of Chicago, which has an 8 month heating season, an increase in the tree cover by 11% in an urban block would potentially reduce annual energy demands for house heating and cooling by up to 3.8% [101]. Here, the wind shielding effect was most important to reduce energy needs.

A field experiment in Sacramento where 16 trees were planted to the southeast and southwest of two houses and where indoor and outdoor temperatures, roof and wall surface temperatures, wind speed, and air-conditioning electricity use was measured during June–October resulted in an estimated cooling energy savings of 29% during the season for the two houses. Peak demand savings for the same houses were estimated to be 47% and 26%, respectively [102].

Roof and wall greening have the potential to decrease energy consumption of buildings for cooling buildings by improving insulation and reducing thermal loadings. In hot climates, reduced heat load during summertime is of significant importance. As an example, cooling load for a two-story nursery school building in Athens (Greece) was reduced by 6–49% during the summer for the whole building and by 12–87% for the top floor [103]. A study of an eight-story building in Madrid in Spain identified that a green roof reduced the building energy demand by 1% annually and 6% during the summer [104]. In a peak hot weather situation, the cooling load was reduced by 25% for the floor below the green roof. This indicates that green roofs have the greatest effect on energy consumption for buildings with a large roof area compared to the height of the wall, such as single-family houses, warehouses, and big box stores.

This way of reducing energy by tree planting, roof and wall greening has the spin-off of greening the city. However, potential conflicts with modern “green architecture” [105, 106] may also arise as, for instance, trees, wall and roof greening may be an obstacle for collecting solar energy by photovoltaics. This is certainly an important area for cooperation between architects and landscape architects to solve emerging conflicts and create synergies in the integrated design of the green infrastructure and green architecture.

Finally, vegetation, and in particular trees and shrubs have also the capacity to capture air pollutants [42]. More green may offset to some degree the potential increase of air pollution caused by intensification of the heat island. However, as with carbon sequestration, few data indicate that this effect will be quite limited [42]. Therefore, planting more trees and shrubs is not to be considered as an effective means solve the problem of emission of carbon dioxide and air pollutants.

Vulnerability of the Green Infrastructure to Climate Change Impacts

Green infrastructure has a series of potentials in terms of climate change mitigation and adaptation in urban areas. At the same time, green infrastructure is vulnerable to the impacts of climate change. Plantings, habitats, and species are challenged by higher temperatures and changing precipitation patterns, and the thermal benefits of evapotranspiration rely on lush vegetation. Furthermore, as Wilby and Perry [107] state in a review, there may be increased competition from exotic species, disease and pests may spread, and sea level rise can threaten rare coastal habitats: “Earlier springs, longer frost-free seasons, and reduced snowfall could further affect the dates of egg-laying, as well as the emergence, first flowering and health of leafing or flowering plants. Small birds and naturalized species could thrive in the warmer winters associated with the combined effect of regional climate change and enhanced urban heat island.”

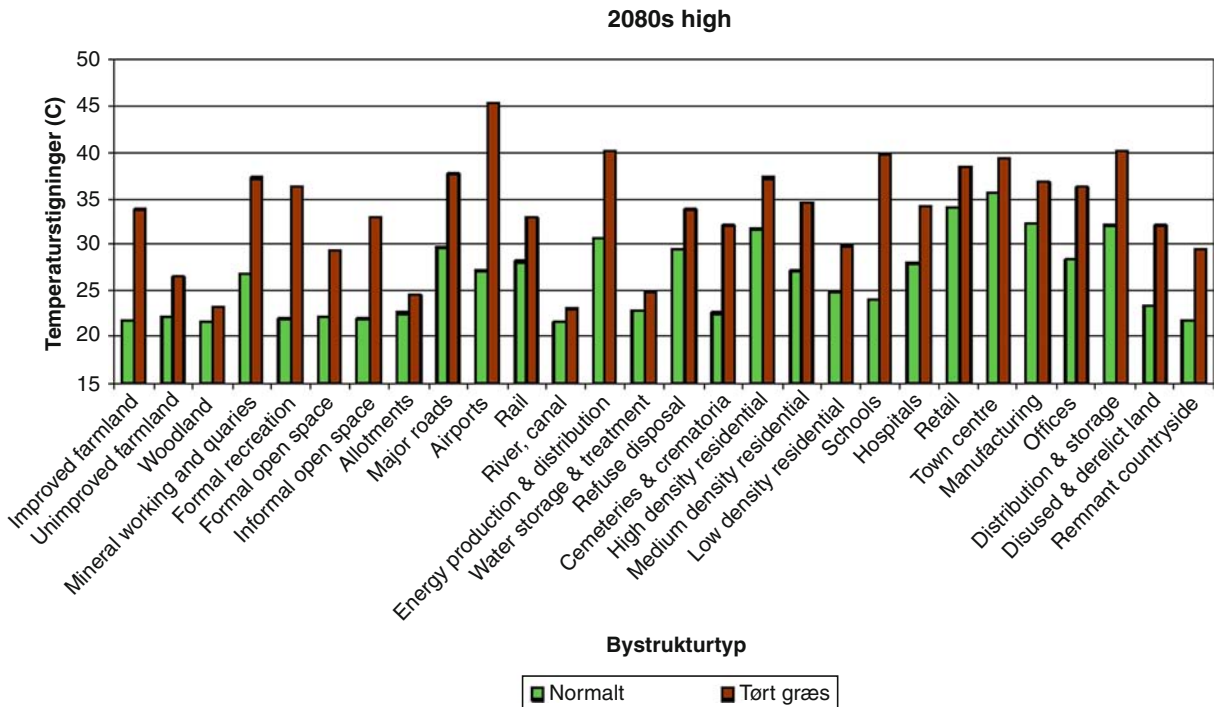
However, the database is very limited at the moment to establish, with any certainty, the impacts of climate change on the urban green infrastructure and to its ecosystem services. In the following, information from the ASCCUE project will be presented to discuss

potential consequences of climate change on the capacity of green space to mitigate the heat-island effect.

Based on the climate scenarios used in the ASCCUE project, it was estimated that periods without water content available to plants in the upper 30 cm of the soil layer would increase from currently less than 1 month during summertime to up to 4–5 month in the inner city of Manchester (UK) in the high-emissions scenario. As a consequence, grassland would dry out and lose its capacity to evapotranspire water. Certainly, reliability of these estimates is low, in particular, as regional soil maps had to be used due to lack of better information on urban soils, which are known to be highly variable [108]. Nevertheless, the figures indicate the potential challenges that may arise for management of the green infrastructure under a changed climate.

Temperature differences with and without the provision of evapotranspiring grassland for 29 urban morphology types in Manchester are shown in Fig. 11. On a hot summer day, surface temperatures would be up to 70% higher on dried-out vegetation than on lush vegetation. In particular, areas with large expanses of lawn such as public parks, playing fields, and also schools would suffer. This points to the increasing need for proper water management to ensure well-growing vegetation or the choice of drought-tolerant vegetation. As concerns the former, mechanisms of water storage for irrigation during times when there is abundant rainfall need to be conceived. Reducing the amount of lawns in favor of trees and shrubs, which can access the water in deeper layers of soil with their roots, should also be considered where possible. However, trees require sufficient space for their roots.

Taking street trees as an example, water stress already occurs as a result of limited soil volume and small rooting space in planting pits combined with soil compaction and generally limited stormwater infiltration in built-up urban areas [109, 110]. Longer dry periods and rising temperatures will increase water stress and result in street trees dying if no proper management regime is installed. For instance, studies from Copenhagen (Denmark) indicated that larger planting pits in combination with improved soil substrates and irrigation can significantly improve growth conditions for street trees [109]. However, taking water for irrigation from the water supply system will become



Green Infrastructure and Climate Change. Figure 11

Surface temperatures on a hot summer day in a 2080s high scenario with lush vegetation and dried-out vegetation ([120] with permission from author)

increasingly problematic under climate change. Another option is to combine irrigation needs with stormwater management, for example by discharging stormwater runoff from pavements to street trees. Such systems are, for example currently tested in the City of Stockholm, Sweden [111].

With changing precipitation patterns, higher temperatures, and later spring frost, the appropriate selection of species is an equally important approach to reduce the vulnerability of green infrastructure to the impacts of climate change. To address the challenge and to raise a discussion on the issue a Climate-Species-Matrix was developed for urban trees in Central Europe [110]. The study identified current regions analogous to the expected future climate in cities in Central Europe (i.e., less than 500 mm precipitation per year and average minimum temperatures between -17.8°C and -23.3°C). It resulted in a list of species with an increased focus on species originating from continental Central Asia and the continental northeastern parts of North America. Here, native species have evolved

under conditions that may be similar to those experienced in urban areas in North and Central Europe in the future.

Future Directions

Climate change will imply significant consequences for the urban climate and the urban hydrology. The urban heat-island effect will increase if no action is taken. But by increasing the provision of urban green space, the temperature rise resulting from global warming can be mitigated. More intensive rain storms resulting from climate change will challenge the capacity of existing urban drainage systems and increase the risk of flooding. Implementing additional landscape-based stormwater detention, infiltration, evapotranspiration, and conveyance measures can compensate for changing precipitation patterns. By promoting urban green infrastructure, a network of green spaces can be implemented in the city to mitigate climate change and to adapt cities to the impact of climate change. A green infrastructure, which is multi-scale,

multifunctional, interconnected, and reflecting hydrology, is a major generator of sustainable urban form. Yet, the green infrastructure needs to be robust to cope with climate change. This has implications for the design of green spaces, including preparation of planting sites and species selection.

Landscape design to face climate change is an emerging concept, which can assist in sustaining cities in an increasingly urbanizing global context. It calls for continuous loops of collaboration, knowledge exchange, experiments, and learning by doing.

A recent symposium with 40 leading European experts [112] led to the development of a framework for research on urban green space. A total of 35 research questions were specified regarding the physicality, experience, valuation, management, and governance of urban green space. At least five of the research questions were directly related to the future role and capacity of urban green infrastructure in the light of climate change. These were primarily related to the physicality of urban green space and included the desired documentation of the “*direct and indirect effects of the climate changes on urban green spaces and how these changes impact people’s well-being in urban areas*,” “*how resilient current green space designs are to climate change and how resilience can be improved*,” and “*how the resilience and adaptability of urban areas can be enhanced to future economic, housing and environmental demands through appropriate design and management of urban green spaces*.” Additionally, 15 research questions were related to aspects of ecological functions, the public goods, and market benefits of urban green infrastructure, interdisciplinarity, and the management and governance of green areas, all addressing appropriate next steps in the exploration, analysis, and understanding of the potential of urban green infrastructure as a tool to face the global challenges of climate change and urbanization.

Adoption of a multidisciplinary approach to the development of the multifunctional green infrastructure is needed, including not only landscape architects and urban planners but also traffic engineers, hydrologists, biologists, sociologists, and economists. Importantly, urban green infrastructure calls for a participatory, socially inclusive approach to its planning and implementation as it will go across public and private land and affect all citizens in different ways. This challenges the sector-based distribution of work and responsibility

areas currently prevalent in most public administrations. Multi-scale approach also involves interinstitutional collaboration and close cooperation between local, regional, and national authorities. Research in the field is emerging in Australia [113], the Netherlands [114], and the USA [115]. The examples from Copenhagen and Munich in this paper have shown that there is increasing potential for such an approach as disciplines realize the limitations of sectoral approaches.

Bibliography

Primary Literature

1. IPCC (Intergovernmental Panel on Climate Change) (2007) Climate change 2007: synthesis report. Contribution of working groups I, II and III to the fourth assessment report of the intergovernmental panel on climate change [Core Writing Team, Pachauri RK, Reisinger A (eds)]. IPCC, Geneva, Switzerland, p 104
2. Parry M, Carter T (1998) Climate impact and adaptation assessment. Earthscan, London
3. MEA (Millennium Ecosystem Assessment) (2005) Ecosystems and human well-being: synthesis. Island, London. <http://www.millenniumassessment.org/documents/document.356.aspx.pdf>. Accessed 4 Nov 2010
4. Alberti M (2008) Advances in urban ecology. Springer, New York
5. Pauleit S, Breuste JH (2011) Land use and surface cover as urban ecological indicators. In: Niemelä J (ed) Handbook of urban ecology. Oxford University Press, Oxford, pp 19–30
6. US Geological Survey (2010) The water cycle: evapotranspiration. <http://ga.water.usgs.gov/edu/watercycleevapotranspiration.html>. Accessed 6 May 2010
7. Lyr H, Fiedler H-J, Tranquillini W (1992) Physiologie und Ökologie der Gehölze. G. Fischer Verlag, Jena & Stuttgart
8. Benedict MA, McMahon ET (2006) Green infrastructure: linking landscapes and communities. Island, Washington, DC
9. Ahern J (2007) Green infrastructure for cities: the spatial dimension. In: Brown P, Novotny V (eds) Cities of the future: towards integrated sustainable water and landscape management. IWA Publishers, London, pp 267–283
10. Pauleit S, Liu L, Ahern J, Kazmierczak A (2011) Multifunctional green infrastructure planning to promote ecological services in the city. In: Niemelä J (ed) Handbook of urban ecology. Oxford University Press, Oxford, pp 272–285
11. Holling CS (1973) Resilience and stability of ecological systems. *Annu Rev Ecol Syst* 4:1–23
12. Gunderson LH, Holling CS, Light SS (1995) Barriers broken and bridges built: a synthesis. In: Gunderson LH (ed) Barriers and bridges to the renewal of ecosystems and institutions. Columbia University Press, New York, pp 489–532
13. IPCC (Intergovernmental Panel on Climate Change) (2001) Climate change 2001. Overview of impacts, adaptation, and vulnerability to climate change. Working group II contribution

- to the third assessment report of the intergovernmental panel on climate change. IPCC, Geneva, Switzerland, p 89
14. United Nations (2008) World urbanization prospects: the 2007 revision. United Nations Department of Economic and Social Affairs/Population Division, New York
 15. Angel S, Sheppard SC, Civco DL (2005) The dynamics of global urban expansion. Transport and urban development department. The World Bank, Washington, DC
 16. O'Meara M (1999) Reinventing cities for people and the planet, Worldwatch Paper, 147. Worldwatch Institute, Washington, DC, p 94
 17. Oke TR (1997) Urban climates and global change. In: Perry A, Thompson R (eds) Applied climatology: principles and practice. Routledge, London, pp 273–287
 18. Satterthwaite D (2008) Cities' contribution to global warming: notes on the allocation of greenhouse gas emissions. Environment and Urbanization 20:539–549
 19. Hunt A, Watkiss P (2007) Literature review on climate change impacts on urban city centres: initial findings. OECD Working Paper ENV/EPOC/GSP(2007)10/FINAL. <http://www.oecd.org/dataoecd/52/50/39760257.pdf>. Accessed 31 May 2010
 20. Blanco H, Alberti M (2009) Building capacity to adapt to climate change through planning. In: Blanco H, Alberti M, Forsyth A, Krizek KJ, Rodriguez DA, Talen E, Ellis C (eds) Hot, congested, crowded and diverse: emerging research agendas in planning. Prog Plan 71:158–169
 21. Newman P, Kenworthy JR (1989) Sustainability and cities: overcoming automobile dependence. Island, Washington, DC
 22. EEA (European Environment Agency) (2006) Urban sprawl in Europe. The ignored challenge. EEA Report No 10/2006. Office for Official Publications of the European Communities, Luxembourg
 23. Nilsson K, Nielsen TS, Pauleit S (2008) Integrated European research on sustainable development and peri-urban landuse relationships. Urbanistica 138:106–109
 24. Haase D (2011) Processes and impacts of urban shrinkage and response by planning. In: Encyclopedia of sustainability science and technology. Springer, Berlin
 25. Wilby RL (2007) A review of climate change impacts on the built environment. Built Environ 33(1):31–45
 26. Crichton D (2001) The implications of climate change for the insurance industry – an update and outlook to 2020. BRE, Watford
 27. SoU (Swedish Commission on Climate and Vulnerability) (2007) Sweden facing climate change – threats and opportunities. Swedish Government Official Reports SOU 2007:60. Swedish Commission on Climate and Vulnerability, Stockholm
 28. EEA (European Environment Agency) (2008) Impacts of Europe's changing climate – 2008 indicator-based assessment. EEA Report No 4/2008. Office for Official Publications of the European Communities, Luxembourg
 29. McGranahan G, Balk D, Anderson B (2007) The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. Environment and Urbanization 19(1):17–37
 30. O'Brien K, Eriksen S, Sygna L, Nygaard L (2007) Why different interpretations of vulnerability matter in climate change discourses. Clim Policy 7:73–88
 31. Kelly PM, Adger WN (2000) Theory and practice in assessing vulnerability to climate change and facilitating adaptation. Clim Change 47:325–352
 32. Bridgeman H, Warner R, Dodson J (1995) Urban biophysical environments. Oxford University Press, Oxford
 33. Lindley SJ, Handley JF, Theuray N, Peet E, Mcevoy D (2006) Adaptation strategies for climate change in the urban environment: assessing climate change related risk in UK urban areas. J Risk Res 9(5):543–568
 34. O'Brien K, Sygna L, Haugen JE (2004) Vulnerable or resilient a multi-scale assessment of climate impacts and vulnerability in Norway. Clim Change 64:193–225
 35. Mehrotra S, Natenzon CE, Omojola A, Folorunsho R, Gilbride J, Rosenzweig C (2009) Framework for city climate risk assessment. In: Fifth urban research symposium 2009, Marseille, June 28–30, 2009. <http://siteresources.worldbank.org/INTURBANDEVELOPMENT/Resources/336387-1256566800920/6505269-1268260567624/Rosenzweig.pdf>. Accessed 31 May 2010
 36. Blanco H, Alberti M, Forsyth A, Krizek KJ, Rodriguez DA, Talen E, Ellis C (2009) Hot, congested, crowded and diverse: emerging research agendas in planning. Prog Plan 71: 153–205
 37. Chiesura A (2004) The role of urban parks for the sustainable city. Landscape Urban Plan 68:129–138
 38. Tyrväinen L, Pauleit S, Seeland K, de Vries S (2005) Benefits and uses of urban forests and trees: a European perspective. In: Konijnendijk CC, Nilsson K, Randrup TB, Schipperijn J (eds) Urban forests and trees in Europe – a reference book. Springer, Berlin, pp 81–114
 39. Tzoulas K, Korpela K, Venn S, Yli-Pelkonen V, Kazmierczak A, Niemela J, James P (2007) Promoting ecosystem and human health in urban areas using green infrastructure: a literature review. Landscape Urban Plan 81(3):167–178
 40. Whitford V, Ennos AR, Handley JF (2001) "City form and natural process" – indicators for the ecological performance of urban areas and their application to Merseyside, UK. Landscape Urban Plan 57(2):91–103
 41. Gartland L (2008) Heat islands. Understanding and mitigating heat islands in urban areas. Earthscan, London
 42. Nowak DJ (2002) The effects of urban forests on the physical environment. In: Randrup TB, Konijnendijk CC, Christophersen T, Nilsson K (eds) COST action E12 urban forests and urban trees. Proceedings No. 1. Office for Official Publications of the European Communities, Luxembourg, pp 22–42
 43. Girardet H (2004) Cities people planet: liveable cities for a sustainable world. Wiley, Chichester
 44. Landsberg HE (1981) The urban climate. Academic, New York
 45. Oke TR (1987) Boundary layer climates, 2nd edn. Routledge, London, New York
 46. Akbari H, Pomerantz M, Taha H (2001) Cool surfaces and shade trees to reduce energy use and improve air quality in urban areas. Sol Energy 70(3):295–310

47. Robine JM, Cheung SL, Le Roy S, Van Oyen H, Herrmann SR (2008) Report on excess mortality in Europe during summer 2003. EU Community Action Programme for Public Health, Grant Agreement 2005114. http://ec.europa.eu/health/ph_projects/2005/action1/docs/action1_2005_a2_15_en.pdf. Accessed 27 Jan 10
48. USEPA (United States Environmental Protection Agency) (2001) Inside the greenhouse: a state and local resource on global warming. USEPA (United States Environmental Protection Agency), Washington, DC
49. Robel F, Hoffmann U, Riekert A (1978) Daten und Aussagen zum Stadtklima von Stuttgart auf der Grundlage der Infrarot-Thermographie. Beiträge zur Stadtentwicklung Nr. 15. Landeshauptstadt Stuttgart
50. Taha H (1997) Urban climates and heat islands: albedo, evapotranspiration, and anthropogenic heat. *Energy Build* 25(2): 99–103
51. von Stülpnagel A (1987) Klimatische Veränderungen in Ballungsgebieten unter besonderer Berücksichtigung der Ausgleichswirkung von Grünflächen, dargestellt am Beispiel von Berlin (West). Unpublished PhD thesis, TU Berlin, Berlin
52. Eliasson I, Upmanis H (2000) Nocturnal air flow from urban parks – implications for city ventilation. *Theor Appl Climatol* 66:95–107
53. Handley J (2006) Adaptation strategies for climate change in the urban environment (ASCCUE). In: Walsh CL, Hall JW, Street RB, Blanksby J, Cassar M, Ekins P, Glendinning S, Goodess CM, Handley J, Noland R, Watson SJ (eds) Building knowledge for a changing climate: collaborative research to understand and adapt to the impacts of climate change on infrastructure, the built environment and utilities. Newcastle University, March 2007, pp 44–53. http://www.ukcip.org.uk/images/stories/Pub_pdfs/BKCC-Results.pdf. Accessed 11 April 2010
54. Gill S, Handley J, Ennos R, Pauleit S (2007) Adapting cities for climate change: the role of the green infrastructure. *Built Environ* 30(1):97–115
55. Gill S, Handley J, Pauleit S, Ennos R, Theuray N, Lindley S (2008) Characterising the urban environment of UK cities and towns: a template for landscape planning in a changing climate. *Landscape Urban Plan* 87:210–222
56. Matzarakis A, Mayer H, Iziomon M (1999) Applications of a universal thermal index: physiological equivalent temperature. *Int J Biometeorol* 43:76–84
57. Dousset B, Gourmelon F, Laaidi K, Zeghnoun A, Giraudet E, Bretin P, Vandentorren S (2009) Satellite monitoring of summertime heat waves in the Paris metropolitan area. In: The seventh international conference on urban climate, 29 June – 3 July 2009, Yokohama, Japan. http://www.ide.titech.ac.jp/~icuc7/extended_abstracts/pdf/384388-1-090518140731-002.pdf. Accessed 31 May 2010
58. BETWIXT (2005) Built Environment: weather scenarios for investigation of impacts and extremes. Daily time-series output and figures from the CRU weather generator [online]. http://www.cru.uea.ac.uk/cru/projects/betwixt/cruwg_daily/. Accessed 30 June 2008
59. Watts M, Goodess CM, Jones PD (2004) The CRU daily weather generator. Climatic Research Unit, University of East Anglia, Norwich
60. Burkhardt I, Dietrich R, Hoffmann H, Leschnar J, Lohmann K, Schoder F, Schultz A (2008) Urbane Wälder. Bundesamt für Naturschutz (eds), Naturschutz und Biologische Vielfalt, 63. Bonn-Bad Godesberg, pp 214
61. Pauleit S, Golding Y, Ennos R (2005) Modeling the environmental impacts of urban land use and land cover change – a study in Merseyside, UK. *Landscape Urban Plan* 71(2–4):295–310
62. Perry T, Nawaz R (2008) An investigation into the extent and impacts of hard surfacing of domestic gardens in an area of Leeds, United Kingdom. *Landscape Urban Plan* 86:1–13
63. Madsen H, Arnbjerg-Nielsen K, Mikkelsen PS (2009) Update of regional intensity–duration–frequency curves in Denmark: tendency towards increased storm intensities. *Atmos Res* 92(3):343–349
64. DMI (Danish Meteorological Institute) (2007) Klimaet i Danmark i 2100 i forhold til 1990 for A2- og B2-scenariene. http://www.dmi.dk/dmi/index/klima/fremtidens_klima-2/aendringer_i_danmark.htm. Accessed 11 April 2010 (in Danish)
65. Arnbjerg-Nielsen K (2008) Forventede ændringer i ekstremregn som følge af klimaændringer. Spildevandskomiteen, Skrift nr. 29, IDA Spildevandskomiteen, Danish Society of Engineers, Copenhagen (in Danish)
66. Rambøll (2008) Kommunernes investeringsbehov i forbindelse med klimatilpasning og veje. Local Government Denmark, Copenhagen. <http://www.ramboll-management.dk/news/~media/Images/RM/RM%20DK%20and%20RM%20Group/PDF/Publications/2009/KommunernesInvesteringsbehovIforbindelseMedKlimatilpasningOgVej.ashx>. Accessed 11 April 2010 (in Danish)
67. Butler D, Parkinson J (1997) Towards sustainable urban drainage. *Water Sci Technol* 35(9):53–63
68. Chocat B, Ashley R, Marsalek J, Matos MR, Rauch W, Schilling W, Urbonas B (2007) Toward the sustainable management of urban storm-water. *Indoor Built Environ* 16(3):273–285
69. Marsalek J, Chocat B (2002) International report: stormwater management. *Water Sci Technol* 46(6–7):1–17
70. Coombes PJ, Argue JR, Kuczera G (2000) Figtree place: a case study in water sensitive urban development (WSUD). *Urban Water* 1(4):335–343
71. Wong THF (2006) An overview of water sensitive urban design practices in Australia. *Water Pract Technol* 1(1)
72. Geiger WF, Dreiseitl H (1995) Neue Wege für das Regenwasser. R. Oldenbourg Verlag, Munich
73. Beenen T, Boogaard FC (2007) Lessons from 10 years storm water infiltration in the Dutch Delta. In: Proceedings of the 6th international conference on sustainable techniques and strategies in urban water management, Novatech 2007, June 25–28 2007, Lyon, pp 1139–1146
74. Scholz M (2006) Best management practice: a sustainable urban drainage system management case study. *Water Int* 31(3):310–319

75. Stahre P (2006) Sustainability in urban storm drainage: planning and examples. Svenskt Vatten, Stockholm
76. Bengtsson L (2002) Avrinning från gröna tak (runoff from greenroofs). Vatten 58:245–250
77. Villarreal EL, Semadeni-Davies A, Bengtsson L (2004) Inner city stormwater control using a combination of best management practices. Ecol Eng 22:279–298
78. Beneke G (2003) Regenwasser in Stadt und Landschaft – Vom Stückwerk zur Raumentwicklung – Plädoyer für eine Umorientierung. Beiträge zur Räumlichen Planung – Schriftenreihe des Fachbereichs Landschaftsarchitektur und Umweltentwicklung der Universität Hannover Heft 70
79. Fryd O, Backhaus A, Jeppesen J, Ingvertsen ST, Birch H, Bergman M, Petersen TEP, Fratini C (2009) Koblede afkoblinger. http://www.2bg.dk/Internal_Workshop/2009-12-03-KE/2BG_HarrestrupAa_Booklet_web.pdf. Accessed 11 April 2010 (in Danish)
80. Turenscape (2004) Yongning river park, Taizhou city. <http://www.turenscape.com/English/projects/projectphp?id=323>. Accessed 6 June 2010
81. Stokman A, von Seggern H, Rabe S, Schmidt A, Werne J, Zeller S (2008) Wasseratlas: Wasserland-Topologien für die Hamburger Elbinsel. Jovis Verlag, Berlin
82. Mathur A, da Cunha D (2009) SOAK – Mumbai in an estuary. Rupa, New Delhi
83. van Nieuwenhuijze L (2006) (Hoog)water als uitdaging Meervoudig gebruik van de dijk en het buitendijkse gebied: wie durft? Report developed by H + N + S Landscape Architects, Utrecht. (in Dutch). <http://www.hns-land.nl/images/stories/Publicaties/hoogwaterstrategie72dpi.pdf>. Accessed 6 June 2010
84. Dreiseitl H (2009) Bishan park, Singapore. <http://www.dreiseitl.net/index.php?id=525&lang=en&choice=58&ansicht=text>. Accessed 05 June 2010
85. Mathur A, de Cunha D (2001) Mississippi floods: designing a shifting landscape. Yale University Press, New Haven
86. City of Los Angeles (2007) Los Angeles river revitalization master plan. http://www.larivermp.org/CommunityOutreach/masterplan_download.htm. Accessed 11 April 2010
87. Rinaldi BM (2007) Landscapes of metropolitan hedonism. The cheonggyecheon linear park in Seoul. J Landscape Archit 2007:60–73
88. Nyhuus S (2005) Oslo. In: Werquin AC, Duhem B, Lindholm G, Oppermann B, Pauleit S, Tjallingii S (eds) Green structure and urban planning. Final report. COST Action C11, European Commission, Brussels, pp 184–191. <http://www.greenstructureplanning.eu/COSTC11-book/>. Accessed 11 April 2010
89. Attwell K (2005) Green planning as a prerequisite for urban development in Aarhus, Denmark. In: Werquin AC, Duhem B, Lindholm G, Oppermann B, Pauleit S, Tjallingii S (eds) Green structure and urban planning. Final report. COST Action C11, European Commission, Brussels, pp 345–351. <http://www.greenstructureplanning.eu/COSTC11-book/>. Accessed 11 April 2010
90. Oppermann B (2005) Redesign of the river Isar in Munich, Germany. Getting coherent quality for green structures through competitive process design? In: Werquin AC, Duhem B, Lindholm G, Oppermann B, Pauleit S, Tjallingii S (eds) Green structure and urban planning. Final report. COST Action C11, European Commission, Brussels, pp 372–378. <http://www.greenstructureplanning.eu/COSTC11-book/>. Accessed 11 April 2010
91. City of Melbourne (2009) Zero net emissions by 2020 – update 2008. http://www.melbourne.vic.gov.au/Environment/WhatCouncilisDoing/Documents/zero_net_e-missions_2020.pdf. Accessed 11 April 2010
92. Kågeson P (2005) Reducing CO₂ emissions from new cars. European federation for transport and environment, Brussels, p 10. <http://www.gronabilister.se/grafik/dynamiskapdf/20050124210807.pdf>. Accessed 11 April 2010
93. New York City (2010) Million trees NYC. New York City Department of Parks & Recreation and New York Restoration Project. http://www.milliontreesnyc.org/html/urban_forest/urban_forest_benefits.shtml. Accessed 11 April 2010
94. Morse SC (1978) Trees in the town environment. J Arboric 4:1–6
95. Foster R, Blaine J (1978) Urban trees survival: trees in the sidewalk. J Arboric 4:14–17
96. Pauleit S, Jones N, Garcia-Marin G, Garcia-Valdecantos J-L, Rivière LM, Vidal-Beaudet L, Bodson M, Randrup TB (2002) Tree establishment practice in towns and cities – results from a European survey. Urban Forestry and Urban Greening 1(2):83–96
97. Bradshaw A, Hunt B, Walmsley T (1995) Trees in the urban landscape. Principles and Practice, Spon, London
98. Nowak DJ, Stevens JC, Sisinni SM, Luley CJ (2002) Effects of urban tree management and species selection on atmospheric carbon dioxide. J Arboric 28(3):113–122
99. Heisler G (1986) Energy savings with trees. J Arboric 12(5):113–125
100. Huang J, Ritschard R, Sampson N, Taha H (1992) The benefits of urban trees. In: Akbari H, Davis S, Dorsano S, Huang J, Winnett S (eds) Cooling our communities. US Environmental Protection Agency, Washington, DC, pp 27–42
101. Jo HK, McPherson EG (2001) Indirect carbon reduction by residential vegetation and planting strategies in Chicago, USA. J Environ Manage 61:165–177
102. Akbari H, Kurn DM, Bretz SE, Hanford JW (1997) Peak power and cooling energy savings of shade trees. Energy Build 25:139–148
103. Santamouris M, Pavloua C, Doukasa P, Mihalakakoub G, Synnefaa A, Hatzibiroa A, Patargias P (2007) Investigating and analysing the energy and environmental performance of an experimental green roof system installed in a nursery school building in Athens, Greece. Energy 32(9):1781–1788
104. Saiz S, Kennedy C, Bass B, Pressnail K (2006) Comparative life cycle assessment of standard and green roofs. Environ Sci Technol 40:4312–4316

105. Brown DE, Fox M, Pelletier MR (eds) (2001) Sustainable architecture white papers. Earth Pledge Foundation, New York
106. Jodidio P (2009) Green architecture now! Taschen, Cologne
107. Wilby RL, Perry GLW (2006) Climate change, biodiversity and the urban environment: a critical review based on London, UK. *Prog Phys Geogr* 30(1):73–98
108. Sauerwein M (2011) Urban soils – characterization, pollution and relevance in urban ecosystems. In: Niemelä J (ed) *Handbook of urban ecology*. Oxford University Press, Oxford, pp 45–58
109. Bühler O, Nielsen CN, Kristoffersen P (2006) Growth and phenology of established *Tilia cordata* street trees in response to different irrigation regimes. *Arboriculture & Urban Forestry* 32(1):3–9
110. Roloff A, Korn S, Gillner S (2009) The climate-species-matrix to select tree species for urban habitats considering climate change. *Urban Forestry and Urban Greening* 8:295–308
111. Alvm B-M, Bennerseid C (2009) Baumstandortoptimierung und Regenwasserbewirtschaftung – Chancen für ein gemeinsames Vorgehen. In: Dujesiefken D (ed) *Jahrbuch der Baumpflege* 2009. Taspo Fachbuchservice, Braunschweig, pp 70–78
112. James P, Tzoulas K, Adams MD, Annett P, Barber A, Box J, Breuste J, Cooper I, Curwell SR, Elmqvist T, Flood T, Frith M, Gledhill DG, Goode D, Gordon C, Greening KL, Handley J, Harding S, Haworth S, Hesketh F, Home R, Johnston M, Kazmierczak AE, Korpela K, Leeks G, Leeks G, Morley E, Nail S, Niemelä J, Moretti M, Stein N, Pauleit S, Powell JA, Radford KG, Richardson D, Roe MH, Sadler JP, Selman P, Scott AV, Snep R, Stern N, Timmermans W, Ward-Thompson C (2009) Urban green – towards an integrated understanding of greenspace in the built environment. *Urban Forestry and Urban Greening* 8(2):65–76
113. Brown RR, Farrelly MA (2009) Challenges ahead – social and institutional factors influencing sustainable urban stormwater management in Australia. *Water Sci Technol* 59(4):653–660
114. van der Brugge R (2009) Transition dynamics in social-ecological systems: the case of Dutch water management. PhD thesis, Erasmus University, Rotterdam
115. Roy AE, Wenger SJ, Fletcher TD, Walsh CJ, Ladson AR, Shuster WD, Thurston HW, Brown RR (2008) Impediments and solutions to sustainable, watershed-scale urban stormwater management: lessons from Australia and the United States. *Environ Manage* 42(2):344–359
116. Handley J (2007) Planning for climate change. Unpubl. Presentation given at the “Future of Cities” 51st international federation for housing programmes (IFHP), World Congress, Copenhagen, Sept 22–26 2007
117. CEC (Commission of the European Communities) (2007) Green paper on adapting to climate change in Europe – options for EU action. Brussels, 29.6.2007. COM(2007) 354 final
118. Voogt JA (2004) Urban heat islands: hotter cities. ActionBioscience.org, American Institute of Biological Sciences. <http://www.actionbioscience.org/environment/voogt.html>. Accessed 07 May 2010
119. Sieker F (Hrsg) (1998) *Naturnahe Regenwasserbewirtschaftung*, Reihe Stadtökologie, Bd. 1, Analytica Verlagsgesellschaft, Berlin
120. Gill S (2006) Climate change and urban greenspace. Unpublished PhD thesis, School of Environment and Development, University of Manchester, Manchester

Books and Reviews

- Askari H, Davis S, Dorsano S, Huang J, Winnett S (eds) (1992) *Cooling our communities – a guidebook on tree planting and light-colored painting*. US Environmental Protection Agency, Washington, DC
- Woods-Ballard B, Kellagher R, Martin P, Jefferies C, Bray R, Shaffer P, Kellagher R (2007) *The SUDS manual (C697)*. Construction Industry Research & Information Association (CIRIA), London

Green Roof Infrastructures in Urban Areas

MANFRED KÖHLER¹, ANDREW MICHAEL CLEMENTS²

¹University of Applied Sciences, Neubrandenburg, Germany

²Green Roof Greece, Athens, Greece

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Acknowledgment
Bibliography

Glossary

FLL The Landscape Research and Development Society (FLL) nonprofit organization was founded in 1975. Its mission is to research, produce, and disseminate all the various landscape development principles, guidelines, and specifications for the assurance of environmental quality [1].

FBB The Green Infrastructure Association (FBB) is a specialized group that was founded by some members of FLL to focus more specifically on green building. The FBB is the German counterpart

to the American industry association Green roofs for Healthy Cities (GRHC) and one of the founding members of the World Green Infrastructure Network (WGIN). The German Word “Bauwerksbegrünung” has no translation in English – Green infrastructure in the sense of FBB is focused on all forms of urban green.

Extensive green roofs (EGR) Also called natural green roofs, eco-roofs, and oikosteges in Greece are vegetated roof constructions that require low to no maintenance. Drought-adapted plant species are used to create a self-sustaining vegetated surface suitable for nearly all types of buildings. Growing media is usually less than 10 cm, or 3 in. deep [2]. The term “Natural Green Roof” or oikostegi, coined by the authors, denotes a green roof where the main interest is in maximizing natural endemic biodiversity on vegetated roofs. These systems can incorporate irrigated areas with rain or gray water. Natural green roofs are designed, engineered green roof systems which enlist the help of the natural environment. Natural green roofs may also be engineered in such a way to include intensive green roofs on occasion.

Intensive green roofs (IGR) Also known as roof gardens are garden structures on top of buildings and other artificial urban surfaces. In most cases, the growing media is more than 20 cm deep; for trees it can be more than 1 m. IGRs, with structures including lawns, planter boxes, shrubs, and small trees, require the same maintenance as traditional gardens; therefore, some question their environmental value.

Stormwater runoff Rainwater running off impervious surfaces.

Green infrastructure Overall phrase in North America for all types of green roof technology and other types of greenery on buildings, like vertical greening, living walls, and indoor greening systems. Green infrastructure in a wider sense includes photovoltaic technology and rainwater management.

Growing media Engineered substrate for green roof purposes. Green roof substrates typically have low nutrient content and high drainage rates. Typical materials are expanded slate, shale, pumice, or recycled construction material such as broken ceramic tiles.

Green infrastructure Broad term in the North America for all types of green roof technology and other types of greenery on buildings, like vertical greening, living walls, and indoor greening systems. Green infrastructure in a wider sense includes photovoltaic technology and rainwater management. This term also describes the range of materials and technologies used to enhance urban environments. In addition to green roofs, this term also encompasses other related systems such as vegetated facades with climbers or living wall systems, indoor greening systems, rain gardens, photovoltaic systems, and other technologies. Roof greening can be combined with living walls, indoor plants, and ecological landscaping to enhance the built environment. The USEPA refers to structures specifically intended to manage wet weather as green infrastructure.

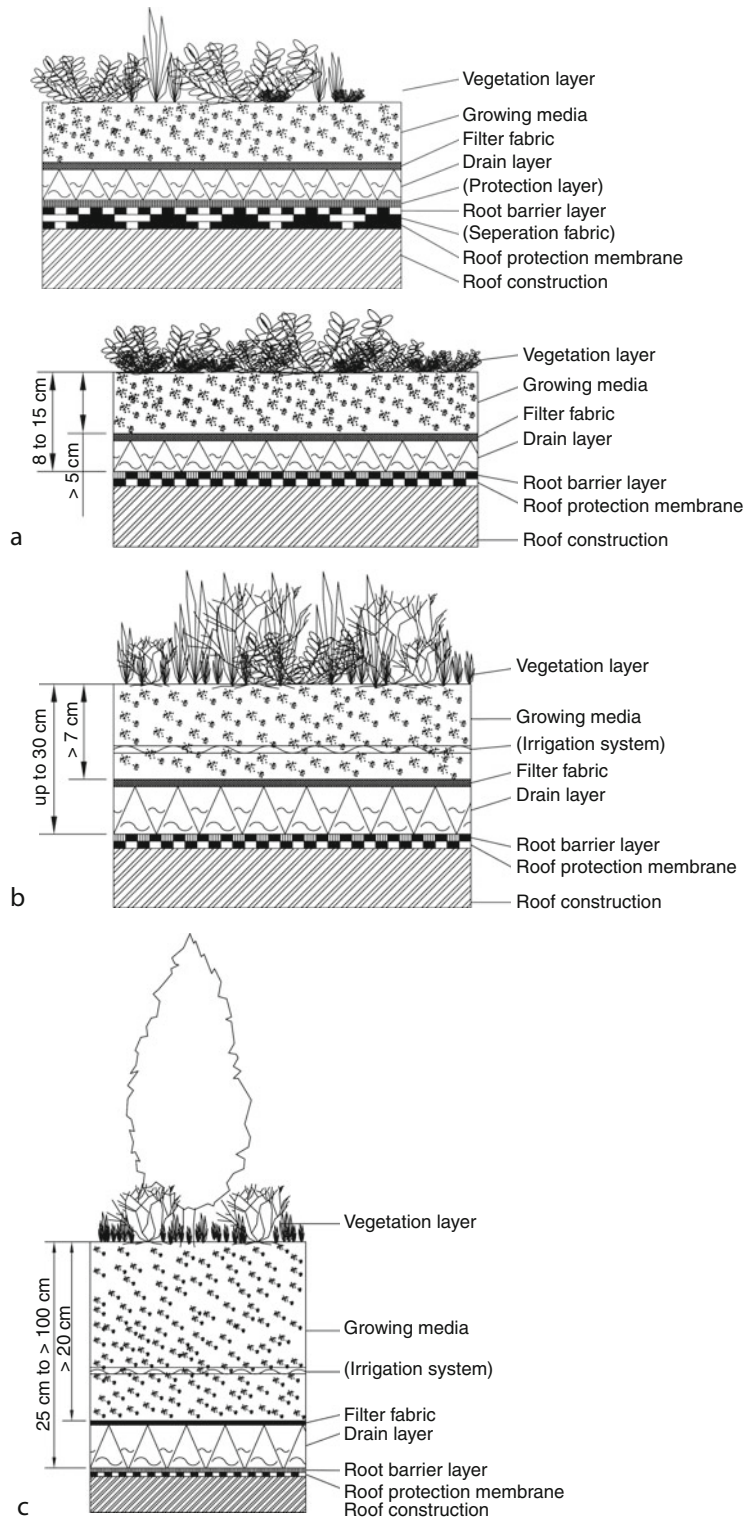
Leadership in Energy and Environmental Design (LEED) This is a US-based rating system by the US Green Building Council (USGBC). Benchmarks focus on energy savings, water efficiency, CO₂ emissions reduction, improved indoor environmental quality, and stewardship. The categories of achievement are silver, gold, or platinum. In Australia, a similar rating system uses “stars.” After an extensive debate about the merits of such certification systems, Germany set one up in 2009. Certification can be an effective type of marketing; however one critique of existing systems is that there is not enough weight placed on vegetation.

Low-impact development (LID) A storm water management and site-design technique to mimic the situation before construction. Water usage, evaporation, cooling, and water storage and drainage are such benefits of green roof infrastructures.

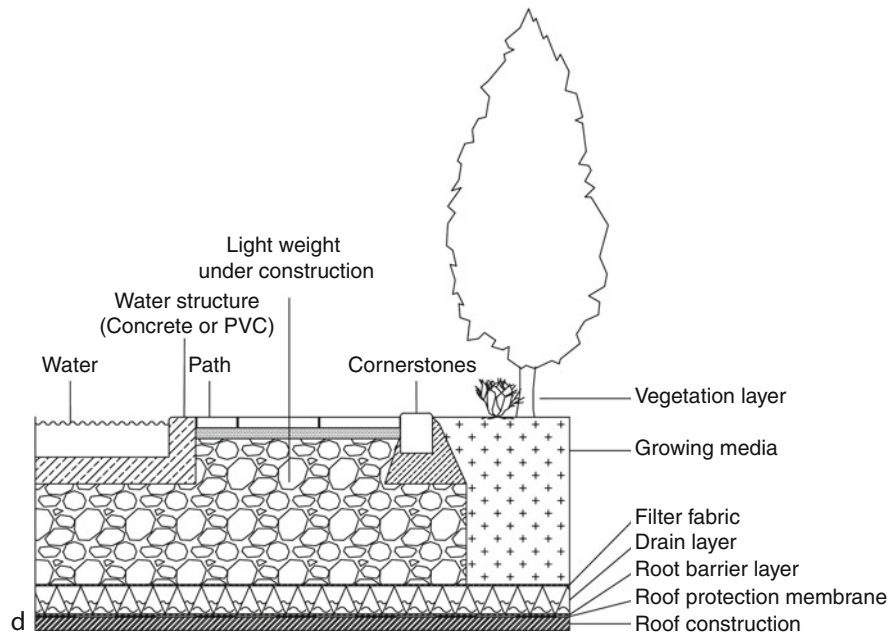
Definition of the Subject and Its Importance

Definition

Green roofs are engineered constructions that include environments suitable for well-adapted plant species. In most cases, these types of roofing have a longer life span than conventional roofing surfaces. The following elements are built on top of the roof structure (from the bottom to the top, see Fig. 1):



Green Roof Infrastructures in Urban Areas. Figure 1 (Continued)



Green Roof Infrastructures in Urban Areas. Figure 1

The first schematic cross section illustrates the principles of green roof structures. (a) Details the one-layer solution without drainage components. This could be difficult in a damp, wet climate. (a1) is a multiple layer solution, typical construction from various suppliers. (b) Semi-extensive means, irrigation is included. (c) Intensive roof garden. (d) Plaza deck construction on garages and similar structures with high load capacity

- The underlying protective layer is made of an impervious material such as bitumen, rubber, polystyrene, or other similarly adequate technical materials, in short, roof protection membranes.
 - Additional, root barrier layers are available to prevent the root penetration of lower layers. These are known as separation fabrics or geotextiles.
 - This is commonly followed by a separate water-retaining layer, which could be a natural porous stone material or an artificial retention mat; in short, this is a drainage layer.
 - On top of this layer, a filter fabric separates the retention layer from the next layer: the growing media.
 - The growing media is, in most cases, a specially mixed lightweight soil material with carefully selected ingredients for storing rainwater. Growing media are mixed for different purposes (for example, extensive green roof growing media differs from roof garden media in nutrient and humus content). Intensive roof garden growing media differs in that in the upper levels there is a higher content of humus and on the lower levels lower humus content.
 - The vegetation layer can range from a shallow layer with mosses and only a few taller plants all the way up to full-blown roof gardens. As such, green roof maintenance requirements can be as little as an annual inspection or as much as is usual for ordinary gardens. The success of the vegetation layer depends on the careful selection of the other green roof layers. For example, if the goal is to plant trees on roof tops, a special combination of all these components is necessary (After [3]).
- The maximum weight of the construction must be calculated carefully. On average, it varies between 40 kg/m^2 (this is about 8.33 lb/ft^2) and can rise upward to about 350 kg/m^2 (71.7 lb/ft^2 , for conversion factors, see [4]) on roof gardens, including the weight associated with water storage. Extensive green roofs should be able to store a minimum of at least 20 L/m^2 of water.

Compared to this load, the weight of all the plant components is relatively small.

The longevity of green roofs depends on whether they can be easily accessed with the basic equipment needed for the success of the project as well as repairs. Maintenance and repairs must be planned carefully. Certain areas of the roof like, edges and places around roof fixtures like skylights or climate control systems can be prone to structural damage. Architects often want to install green roofs on very steep roof inclinations and on very tall buildings. These technical and biological limits challenge green roof professionals. It is the duty of green roof specialists to assess the limits which are set. Uncontrollable aspects of local climate, like wind, temperature, and the intensity of storm events and solar exposure can set limits to what is feasible from an architectural design perspective and ideas about developing living surfaces on roofs and facades.

In Germany [1], two main types of vegetated green roofs are observed:

- Extensive green roofs are large lightweight constructions. Most consist of a layer of under 10 cm deep covered by drought tolerant plant species including mosses, sedum sp, other succulents, wild flowers, aromatic herbs, and a plethora of pioneer species. Commonly, no irrigation is installed, particularly in temperate climate experienced in Central and Northern Europe and North America and only one annual maintenance visit occurs. These structures are possible on nearly all types of flat buildings covered with gravel protection layers; gravel has nearly the same weight as green roof materials but gravel has relatively no environmental value.
- Intensive green roofs, synonymous with roof gardens, are fully maintained with garden structures similar to those on the ground level. The first step in planning such constructions is a careful calculation of the loading capacity.

Recently, [1] a third type, called “simple intensive green” has been developed. These are green roof constructions with a little irrigation on demand, a wider range of plant species, and a little maintenance. Shrubs and garden plant species are suitable for such green roofs. Another difference is the need for maintenance. Extensive green roofs usually only need one inspection

a year. The simple intensive green roofs or roof gardens do need maintenance, and are thus more similar to ground-level gardens.

The phrase “grass or sod roofs” is used to describe extensive green roofs in the Northern European region, while “Living roofs” is the most commonly used term for green roofs in the UK. They distinguish between “green roofs” and “biodiverse green roofs,” which are characterized by a wider range of growing media types at various depths. This creates a wider range of micro-habitats. Furthermore, green roof researchers in the UK also talk about “brown roofs” or roofs with shallow layers of growing media and a sparse vegetation cover, mostly developed for insects and nesting birds. The presence of birds like the Black Redstart on green roofs is a key indicator of success about which people in the UK have become especially emotional. Observations like these help to promote green roofs. In Greece, there is growing interest in another form of extreme natural green roof called an Oikostegi (from the Greek “οικος” meaning home, abode, living quarters, and “στέγη” meaning roof, shelter, protection, den). Oikosteges/οικοστέγες/ are designed for the most extreme built environments on Earth, namely, earthquake zones, which permit only ultra-lightweight building greening systems. In addition, oikosteges are designed to be low to zero maintenance and low to no irrigation in extremely hot, dry, windswept, so-called urban canyons like Athens. This is green infrastructure at its hardest core and at its most extreme.

Installations at prestigious locations such as the Greek Treasury, in Constitution Square, opposite The Greek Parliament in Athens have been acclaimed with rave review by specialists, academia, and the media. Countless species of butterfly, spider, and other insects like honey bees and ladybirds attract numerous species of song birds, which have begun to nest in the center of this sprawling metropolis. This ecosystem has been named the “Meadow of Constitution Square” a touching acknowledgment that nature has begun to establish a foothold in the once barren, windswept capital which was the birthplace of democracy, the rule of law, politics, philosophy, Western medicine, logic, science, the Olympic games to name just a few of the contributions that Greece has made to humanity.

It should be also noted here that Greece has the richest biodiversity in Europe boasting fully one third

of all indigenous species on the European continent. The growth of cities like Athens had all but banished this abundant natural wealth from its home. Oikosteges represent a hopeful sign that Athens and Greece as a whole with its rich natural environment is now reestablishing itself to take the place it deserves as an inspiration and testament to the world of what is possible.

Green roofs are also characterized by their technical structure. The non-ventilated flat or “warm roof” is constructed with the insulation under the waterproofing layer. In contrast, on “inverted roofs” the insulation layer is installed on top of the waterproofing layer. These inverted roofs have received a lot of attention because this layering can protect the waterproofing and also be used to prevent leaks. However, there are also some disadvantages, like, for example, reduced insulation function if the thermal insulation is exposed to water. An open debate about such details of building physics must take place between the roofer and the landscaper executing the green roof. Surface sealing, either with an artificial layer or with liquid sealing components must be selected carefully. There are regional differences in such construction materials. All layers must be qualified as “root resistant” and be secure against aggressive roots. Such layers contain, in many cases, components to prevent root penetration [2]. It should be obvious that these layers should not include poisonous substances. These components could be washed out into the sewer systems or could be a problem for the roof vegetation.

Green roof technology has become internationally renowned in the past few years. The books recommended at the end of this chapter offer a wide range of basic information from various regions in the world.

Introduction

Introduction – Worldwide Historic Roots of Green Roofs

Like many other ground-breaking ideas, the basics of green roofs are simple and some historical examples date back more than 2,000 years [3]. A water barrier layer and an additional layer of growing media make it possible to cover a building with plant material. This coverage protects the building and stabilizes

temperature fluctuations. That is why green roofs can be found worldwide from the hot and dry climates of the African desert zones where there are examples in places like Uganda, to the Northern European zones with long cold frost periods, such as the Iceland sod homes. These prototypes are the predecessors of extensive green roofs.

Roof gardens or intensive Green roofs may have originated from the “hanging gardens” of Semiramis (Babylon). Intensive green roofs can only be built on flat roofs and are quite similar in structure and form to a ground-level garden. In some Mediterranean countries, flat roofs were traditionally used as recreational space for spending time at home in the evening breeze.

Roofs do not necessarily have to be flat in order to be greened; it is also possible to have green sloped or pitched roofs. The grass-covered tombs in the ancient City of Pamukkale (Turkey) are examples of this. Water tanks with a grass roof structure are common in some countries in Southern Europe. On North European pitched rural farmhouse, green roofs are found in many regions. Roof gardens became especially widely distributed [4], creating green islands in urban environments in the 1920s with a peak of interest in the 1970s. Roof gardens as recreational areas on shopping malls and hotels have also become fashionable.

The first emergence of modern green roofs was observed in the 1860s. During this period of industrialization, roofers developed methods of covering their housing tenements with a gravel layer and turf to avoid the risk of fire. Between 1860 and 1920, similar ideas grew in many German Cities. These older green roofs were a cheap way to protect the sensitive layer of tarred cardboard used for roofing at the time [5]. Vegetable gardens were also planted in these first roofs. All these green roofs were erected above a wooden support construction [6].

In the 1920s, several roof garden projects were documented in the literature of the historic garden library; (see the online Databank of the Technical University of Berlin, (<http://freitext.ub.tu-berlin.de/gartenkunst.html>) using the German word “Dachgarten” as a search word. At the time of writing (April 2010), 130 titles are available dating from the beginning of twentieth century concerning green roofs.

Since the beginning of the twentieth century architects, like Le Corbusier and Oscar Niemeyer have

included roof gardens in their designs. Some of these early modern roof gardens still exist today. One of the most famous is the roof garden on the Rockefeller Center. Other examples like the roof gardens at “The Bund” in Shanghai were also built at that time. In the late 1930s, LeCobusier, Oscar Niemeyer, Lucio Costa, and Roberto Burle Marx created roof gardens as part of a workshop at the Ministry of labor and work in Rio de Janeiro, which can still be seen today. The landscape Architect Burle Marx completed nearly 100 roof garden projects mostly in South and North America, but also in some other countries of the world.

Flat and green roofs were the preferred style in “Bauhaus” architecture. Planners in the 1920s were fascinated by the open amenity spaces on the roof tops of modern Cities. In 1929, the storehouse Karstadt at Berlin-Hermannplatz was one such famous building. It was well publicized, and *the* place to stay at that time [4]. In its first year, in 1929, the young Roberto Burle Marx began a 1 year stay with his family in Berlin, studying vegetation in the Botanical Garden and Art. Burle Marx is regarded as being as the landscape Architect of the Bauhaus team, newer publications describe him as this [7, 8]. He is considered to be one of the most innovative landscape architects in the twentieth Century and a pioneer of roof garden technology since the 1940s all over the world. Modern examples are documented as award-winning projects in the USA [9, 10].

Maintenance of these early roof gardens was similar to that of gardens. They were high-quality recreational spaces where the affluent could spend their leisure time. No publications discussing maintenance or issues of ecology are mentioned from this era.

It can be deduced from the painting by Carl Spitzweg, called “The poor poet” 1837, that living in a roof apartment entailed long periods of poor living conditions. In the winter time, it was too cold and in summer it was too hot. This is still the case in countries like Greece. Additional insulation provided by rooftop greenery created better living conditions in the top floor of buildings. This idea, to take advantage of additional insulation while bringing nature back into cities, gained popularity in the years following World War II.

The second emergence of green roof technology in Germany occurred in the late 1970s. A movement of urban environmentalists tried to bring nature back into Cities. A key thinker at that time was the Garden stylist

LeRoy who had ideas about “natural gardens.” People began to reevaluate urban areas [11]. The first ecological studies of green roofs began during this time. Among the pioneers in Germany was, for example, the Architect Minke in Kassel. His focus was the energetic performance of affordable housing. He learned from ancient African architecture and designed simple, modern solutions for “green architecture” in hot climates [12].

Famous artists at that time Friedensreich Hundertwasser and Ben Wargin promoted the concept of more natural cities and especially nature on buildings. But they were not architects. It was a challenge during the following years to materialize these concepts [13].

The Green Roof Movement in Germany

Roof gardens are a well-established technology in many countries. Hotels, holiday resorts, plaza decks, and similar institutions are covered with green roofs in various countries [4]. The main difference between the early stages of the roof garden movement and today is the current focus on ecological function, local plant species selection, and environmentally friendly construction technology.

A water storage and drainage layer can be installed underneath a roof garden, which can then be used as a recreational space for the users, citizens of the entire building. That is the position today. Lots of these older roof gardens failed over the years. Materials knowledge and maintenance techniques have grown over the years. Knowledge has been accumulated over the decades, which means that nowadays green roof technology is vastly superior to earlier attempts.

In the 1970s the so-called “National Nature conservation act” was updated in Germany. Urban areas were included as areas where wildlife should be protected. Local indigenous wildlife was targeted in this act. Naturally occurring urban plants were considered worthy of protection. A new urbanism shifted away from erecting of new neighborhoods on the outskirts of the city, and began to rethink quality of life in urban areas while searching for different places to construct more new apartments. Roofs, which had hitherto been used to dry washing, presented an ideal space to develop green roofs and then develop quality apartments in those buildings.

The Nature Conservation foundation in Berlin “Stiftung Naturschutz”, an interdisciplinary cooperation of architects, planner, and ecologists started to erect such Eco – Roofscapes in Berlin. In many other Germany Cities similar activities started to enhance the quality of life in inner cities using greenery. Community projects such as backyard greening, green facades, and roof greenery were important. The use of local plant species was also a focus. Less emphasis was placed on roof gardens.

The FLL guidelines, which focus on plaza decks and green roofs, integrate both stages. The first FLL guidelines came out in 1982. They addressed all green roof varieties from the extensive to the roof garden types. The first English version was printed in 2002 and then updated in 2008 with the second edition [1].

Disseminating the Ideas of the Modern Green Roof Technology

The idea behind green roofs is basically simple; combine a waterproofing barrier, a water storage and drainage layer, a designed growing media (not typical garden soil), and the initial vegetation – that’s all.

Looking in more detail, there are several concerns related to warranty that must be acknowledged. German homeowners expect their green roofs to be guaranteed for as long as in minimum warranty periods for traditional pitched tile roofs last. If the vegetation cover is not sufficient, wind and storm water erosion can present an increasing problem over the years. An annual inspection should be made to identify any problems. A good quality green roof should have plant coverage of more than 60%. This means that there will be a minimum amount of weeding necessary [14].

The German model has also been successful because of the cooperation between researchers, manufacturers, and planners, under the FLL guideline group. Green roof technology was executed according to certain norms in order to deliver secure, basic, brand-independent solutions. The FLL-approved structures can occasionally be copied, in part, in other countries with similar building codes and climates as Germany. In countries with radically different building conditions and climates, application of such guidelines is disastrous, to say the least. Many aspects of green

roofing which are country or region specific due to the different building codes in different countries and other factors like climate, which affect green roof design and implementation make application of standards established in Germany rather impractical. Despite this, at the European Union level, a CEN Standard working group has started to work on an attempt at a European guideline; this will take time and is probably not practical (please see the discussion about “[Green Roof Incentive Programs](#)” below). The results of these studies may be, in part, used to help form the starting point for attempts at drafting potential national regulations. In the USA, similar activities are also underway to set up standards at the national level. The first standards have been prepared and published [15, 16].

Germany is a mature market where more than 10 million square meters of new green roofs are built each year. In neighboring countries, like France, about 0.5 million square meters are now being built each year. The movement exists in about ten European countries.

Green roof activities started to increase in North America around 2000. Since 2003, the North American Association “Green roofs for healthy Cities” hosts annual conferences to disseminate the green roof concept in North America and the world. In many Asian countries, especially Japan, Korea, and Singapore, many green roofs have been constructed in the last 10 years. In contrast to the Western world, these activities are not organized by private corporations. For example, there are incentive programs to support green roof activities in about 50 cities of Japan. In China, many green roofs are part of new city concepts. In addition, Shanghai pledged to erect about 100,000 m² of new green roofs in 2009. This figure was achieved in the middle of 2009. In Thailand, preliminary green roof studies have been conducted, and more are planned. The majority of the Asian projects are focused on roof gardens to offer additional city recreational space in the mega-metropolitan areas. Extensive green roofs are increasingly used to cover bus shelters and other such structures. The number of hot summer days in sprawling Asian Cities has increased in the last decade. The cooling potential of green roofs as well as reductions of the peak storm water runoff is important benefits for these regions.

In Scandinavia, especially in Norway, grass roofs are a typical construction on old farmhouses. Such structures are not very exportable to other countries for a number of reasons, but they are regionally adapted to this climate. A resurgence of construction of these types of roofs, for summer homes, has experienced a renaissance in recent years. A few cities, like Oslo, and others have been slow to set up general guidelines for green roof projects. They are focused more on architectural design competitions to promote such environmental friendly roofing technologies.

Other countries, like, Spain, Italy, and Greece have also made strides with several new green roof projects in recent years. These consist of both roof gardens and extensive green roofs. In Spain and Italy, about 500,000 m² of green roofs have been installed recently. In the other Mediterranean countries, like Greece, where flat roofs are typical, and the potential benefits of green roofs great, roof greening was seen as unrealistic in the past for various reasons. This changed in 2000 when a team of researchers worked for 5 years to design a Greek roof greening solution, which is adapted to the Greek building code and severe Greek climate. As noted earlier this system is called “oikosteges” and was installed on The Greek Treasury, which is located opposite the Greek Parliament, in the summer of 2008. This installation has been studied twice by the National Metsovio Technical University of Athens by Drs. Rogdakis and Koronaki. These studies concluded that oikostegi significantly effects the thermodynamics of the Treasury leading to appreciable reductions in energy requirement. Photographs of this project can be seen at <http://oikosteges.gr/index.php/photogallery>.

On behalf of the European Association of Green Buildings (EFB), a move was made to set up a European standard for green roof technology in Europe. Such a law would have to focus on the basics of the technology. In this proposal, each country is then requested to adapt this framework law to their specific national needs. For Germany, this would be a step backward since the current FLL guidelines were updated in 2008 [1].

Green roof projects have been registered in more than 40 countries in 2010; see www.worldgreenroof.org. The focus differs among the various climates and regions, but the benefits are quantifiable in all countries.

The main benefits of green roofs are as follows – First of all, green roof structures last longer than

non-green roofs. The longer life span and decreased maintenance cost of the roof is one of the first ways that economic benefit is achieved through implementation of these structures. Different calculation programs demonstrate such benefits. For some examples, refer to Table 1. One current interactive system is the Green roof calculator [17].

Green Roof Incentive Programs

In Germany, in the 1980s, most cities set up programs to encourage urban greening. The basis for these measures was the Environmental nature protection law, updated in the late 1970s, which stated that all regions of Germany had to work on implement urban greening plans. Not only did the regions have to set up specific laws, but the cities also had to set up rules for incentives and tax breaks. An overview of all these different regulations was collected in three rounds of questionnaires distributed by the FBB; see lists on www.fbb.de.

In conclusion, it can be seen that government incentive programs resulted in the development of backyard greenery, green facades, and green roofs in the 1980s. Later on, in 1990, one of the most successful pieces of incentive legislation was implemented whereby homeowners could receive storm water amelioration tax breaks for installing green roofs.

Storm water taxes penalize runoff created by built surfaces, such as roofs, because it burdens the public sewer system. In Germany, such a tax is about €1/m² of built surface. Locally, legislation varies between no rebates for roof greening right up to a 100% storm water tax exemption. On average, in Germany, across the board, this equals about €0.50/m² of green roof [25].

Today, in late 2009, the implementation of green roofs is well established and is included in about one third of all German building plans for permit submission. In Berlin, there is a government department that is responsible for guiding public building projects, monitoring selected key projects, and publishing basic information and guidelines for public housing and public construction sites (see: http://www.stadtentwicklung.berlin.de/bauen/oekologisches_bauen/index.shtml).

Green Roof Policy in Other European Countries In all other European countries, roof greening legislation is more recent than the laws found in Germany. In the

Green Roof Infrastructures in Urban Areas. Table 1 Selected economic studies about the green roof benefits

Type of study	Source	Results	Comments
National survey of extensive green roof cost – benefits in Germany	[18]	Costs: Extensive green structure (m ²): €17, Maintenance 40 Years €17 additional static. €10–44 costs. Benefit: Costs – longer lifespan (up to €37) – reduce rain tax (up to €32) = profit of about: 25€/m ²	Until 2002 no calculation of energy savings had been completed Costs are based on German rates in 2002 which are close to those in 2009
Phillips Eco-Enterprise Center – case study – Warehouse after LEED criteria USA	[19]	Costs of the extensive roof \$78,000 ROI, annual return – \$1,000	Not calculated energy savings, study not complete
US case study for an affordable neighborhood in NYC – USA	[20]	Break even point after 15 years/return on investment	Calculation investment, energy savings etc.
National comparison Germany, Brazil, USA	[21]	Over the lifespan of 90 years, the profit of a roof garden is the highest: plus \$1,200/m ² . Extensive roofs lost value (\$180/m ²) as did Bitumen/gravel roof (minus about \$300/m ²)	Various differences in traditions, technology, and level of salaries
City Study Toronto: (CA)	[22]	If 50% of the roofs of the city of Toronto were extensively greened, there would be annual cost savings: \$37 million	See full study on internet as pdf on the web page of city of Toronto. The results are a summary of the Toronto action plan
Portland City Study – USA	[23]	If flat roofs in a warehouse district of Portland were greened, the cost savings of the reduced storm water entering the sewage system is enough to pay for all these green roofs and reap long-term benefits	These results were integrated City of Portland action plans
Life cycle assessment of Green roof systems in Hong Kong	[24]	In tropical climate, green roofs double the life span of roof materials, reduces energy costs for air conditioning systems. Green roofs are an economically viable solution for investors	Available on web page of University of Hong Kong

1980s, Austria, Switzerland, The Netherlands, and Hungary began to follow Germany's lead. The focus was different and simplified. In Austria, roof gardens were the focus of interest. In Austria, recreational space in Vienna and other large Austrian Cities was at a premium. The focus of the green roof movement in Switzerland is on the biodiversity of extensive green roofs. Similar tendencies, among other trends, can be found today in the UK and Sweden.

Regulations at the national government level are foreseen for other countries by late 2009. Some cities are forging ahead. In the UK, London has set up a green roof city policy for the coming years. Many green roof

related themes can be found on the official city of Greater London web site (<http://search.london.gov.uk/search>, key words living roofs). In June, 2009, an internet search returned about 1,620 related links. Other British cities are behind London, but advocacy groups are forming in many British cities. In Ireland, in 2007, thanks to the Dun Laoghaire Authority, the first seminar to promote green roofs was held.

Worldwide Distribution Over the last 10 years, many other countries have begun to write green roof policies. This has led to the growth of roof greening around the world (Table 2). In the USA, about

1,000,000 m² of green roofs were installed in 2008. Between 2008 and 2009, an increase of 25% of square meter age was calculated by the national trade association (survey, done by Green Roofs for healthy Cities www.greenroofs.org). In Mexico City, in 2007, 10,000 m² and in 2008 8,000 m² green roofs were constructed with incentives given by the metropolitan Mexico City government and the current number of Extensive Green Roofs (EGR) constructed with the support of City administration incentives are now about 25,000 m². (Welcoming remarks Marcelo Ebrad, Mayor of Mexico City, “Congresso Mundial del Azoteas Verdes”, October 7s. 2010) (Tanya Müller and Amenamex personal communication).

In Asia, Singapore started a green roof policy to promote roof gardens in the 1980s. Since about 2002, it has been part of national environmental policy to set up and test extensive green roofs. A special department, called CUGE was founded to promote all types of

environmental friendly building and construction technologies [26].

The other mega cities in Asia are also getting on board. Hong Kong has conducted a city-wide study, which reviewed the potential of the technology [27]. Bangkok is just starting with an academic research group and demonstration roofs. All these cities already have roof gardens for recreational purposes. The newer ones integrate rainwater retention technology.

Seoul, in South Korea, is an example of an Asian city which adapted the idea of “Biotope area factor” from Berlin http://www.stadtentwicklung.berlin.de/umwelt/landschaftsplanung/bff/index_en.shtml). A green roof with a growing media more than 90 cm has a thermal insulation value of 0.7, a green roof with more than 20 cm has 0.6, and a green roof less than 20 cm has 0.5. Also, a planted wall has a value of 0.4. By 2012, the city hopes to vegetate 600 roofs. Fifty percent of the costs are borne by the state. In 2009, more than

Green Roof Infrastructures in Urban Areas. Table 2 Fact Block, selected numbers on green developments around world

Numbers	Locality	Description	Source
7.5%	Berlin	Calculation of the total percentage of green roofs in Berlin City. This is the number calculated counted only for the quarter Berlin – Kreuzberg	B
32%	USA	Growth of the Green roof market between 2007 and 2008	A
8,000 m ²	Mexico City	An area of public green roofs, officially opened by the lord mayor of Mexico City in 2008	C
53,000 m ²	Chicago	Chicago is the number 1 green roof builder in the USA. The total number of green roof projects presently installed are 84	A
100,000 m ²	Shanghai	Green roof area pledged for installation by 2008, the goal was achieved. Since the beginning of the campaign in 2003, 500,000 m ² have now been installed	D
600,000 m ²	Beijing	Total area of green roofs installed in the city as of 2006	G
310,000 m ²	USA/ Canada	The total area of new green roofs installed in 2008	A
1,000,000 m ²	France	The total green roof area expected annually	E
11,000,000 m ²	Germany	This is the rough estimate of new green roofs built each year in Germany, two third of these are extensive	F

A: Green roofs for Healthy City, 2008 Green roof Industry survey

B: Koehler, own survey in 2008, (prepared for publishing at Cities Alive, 2009)

C: Personal communication, Tanya Mueller, Amenamex in July 2009

D: China Daily [30]

E: Adivet

F: Hämmerle [31]

G: Koshimizu and Lee [32]

300 green roofs have been installed equaling 100,000 m² in total. The Women's University in Seoul is a beacon and a prime example of Seoul's earth-covered buildings [28].

Japan supports green roof projects in about 50 cities. Tokyo, for example, provides a tax break of 50% for 5 years after the installation. New buildings larger than 1,000 m² must be offset with at least 20% vegetated surface, which can include the roof surface. There is no general guideline for all Japanese cities [29] (Figs. 2 and 3).

Green roof advocates and activists can be found in nearly every Australian city. They organize workshops and seminars. One of the most prestigious projects using green roof technologies in Australia is on the house of Parliament, in Canberra. It was designed in the late 1970s with a Kentucky blue grass turf roof and many roof gardens, for recreational space for employees. In Sydney, green roof regulations were first instituted in 2008. Other Australian cities are following suit. Main issues in Australia are roof gardens, native plants, and "food on the roofs." The idea of green roofs and water

savings, as well the question to accept dried out vegetation during droughts is on the very beginning.

It is not easy to provide an overview of African activities. There are ancient green roofs in many African countries, like Tanzania and Uganda. Modern projects are beginning to be seen in South Africa, especially at holiday resorts and fashion retail stores.

In summary, there are examples of green roofs in more than 40 countries worldwide. An interest on the ecological benefits of roof greening is growing. Establishing regulations that focus on the main needs of each country will be the next major challenge for local environmentalists.

Longevity of Green Roofs

The reason green roofs last longer than traditional roofs is the additional layering, which blocks direct exposure to solar radiation. Solar exposure causes premature aging of unprotected roof membranes.

The simplest explanation for this benefit is the direct comparison of roof surface temperatures during day-time. Temperatures up to 70°C are possible on a roof



Green Roof Infrastructures in Urban Areas. Figure 2

The grass roof on the Parliament house in Canberra is an semi-extensive green roof, Australia



Green Roof Infrastructures in Urban Areas. Figure 3

Intensive Green roof, one of the roof garden areas for the Australian politicians in the House of Parliament

that has not been greened. Temperatures are reduced on the green roof surface because of the shading that the vegetation layer provides as well as evaporation cooling. On summer nights, the surfaces of black tar roofs are often cooler than green roofs. These extreme temperature ranges experienced by a nonvegetated roof are another reason for their premature aging.

What these effects mean for a building's energy budget depends on the amount of insulation and the thermal flow through the entire roof construction. This can be calculated for the purpose of assessing energy savings. In both hot and cold countries, the thermal insulation properties of green roofs are an important benefit. It can equate to a significant reduction in building energy requirements. Calculations using a data set from a Neubrandenburg roof for 1 year [33] showed that about 100 m³ less heating gas would be required for a single house with normal insulation when a green roof has been installed.

In terms of cooling, impressive thermal insulation properties have been observed in two studies conducted by the National Metsovia University of Athens. They studied the thermodynamics of the oikostegi green roof installed on the Treasury in Constitution Square, in Athens, opposite the Greek

Parliament. They found that total annual heating and cooling energy savings equaled nearly €6,000. They made a number of important observations, which can be summed up by their conclusion "the thermodynamic behavior of the Greek Treasury is being significantly impacted by the oikostegi green roof."

The amount of captured and evaporated water is also an important aspect for determining reductions in summer air conditioning that could be attributed to green roofs. Of course, it is unwise to use savings in energy consumption to irrigate a roof because water is an even more fundamental and important finite resource than energy. Depending on the region, cooling requirement reductions could be more important than heating requirements. The Metsovia Study suggested that it is feasible to approach a reduction of up to 100% in air-conditioning requirements in a building where an oikostegi green roof has been installed.

How to Promote the Idea of "Green Envelopes"

In order to convince clients to install green roofs, practitioners must explain the technology and create demonstration projects to give them a sense of how these structures will look. To convince economists

and legislators, it is important to quantify the economic costs and benefits of green roofs. Most cities start first with incentive programs, for example, the City of Linz (Austria) set up financial support for green projects for a couple of years. Then, after this worked well, they reduced the amount of money provided. Once the technology was established, it took off by itself [34].

In contrast with planting back yards and walls with climbing plants, green roofing technology needs more professional designers and technicians to avoid the potential of waterproofing compromise and other damage to the underlying roofing and construction materials. Berlin has had a long-term program from 1983 to 1994 [13] to incentivize more greenery in the city. This underlying requirement for a professional orientation for green roofs may explain why 466,571 m² green backyards have been constructed but only about 63,575 m² green roofs.

US Green Roof Policy There is incentive legislation in many US cities, like Washington DC, Boston, Baltimore, Portland, and others [35]. The environmental protection agency (EPA) set up state policy programs to support green roofing. One of the earliest of such, in September 2007, focused on the reduction of the storm water runoff using green roofs. The next steps will be to consolidate these steps and enhance the reach of the legislation to include other benefits such as thermal insulation. Policy programs are fundamental, but these should be supported by demonstration projects which are then studied by independent academic research organizations who can quantify the benefits of green roofs. Many American projects are documented by the proud activists who post their activities on the YouTube internet platform. Nearly 3,000 contributions are posted under the title “Green roofs”; the majority of these are from the USA (improved in January 2010). The German term, “Gründächer” only got two hits. This is a mirror of an expanding environmental awareness in the USA. Knowledge of this may help to reinvigorate the green movement in Europe.

One important benefit of green roofs is the reduction of storm water runoff, which has been documented in Europe for nearly 20 years by several research institutes. Similar studies are now underway in about ten

different institutions in the USA such as PennState, and the Lady Bird Johnston Wildflower Center in Austin, Texas. In 2006, the US mayors’ assembly passed a green roof resolution at their annual meeting, at which they stated that “green roofs manage storm water naturally, reduce flooding risk and improves air and water quality.” For only a little extra cost during roof installation, long-term benefits can be achieved.

Green Roof Growing Media and Plant Species Richness

A green roof begins on the final surface provided by the builder whether that be metal, wood, or concrete. The first layer and one of the most fundamental aspects of a green roof is the waterproofing. There are a number of ways that sound waterproofing can be achieved, and this variation is due to different building conditions and codes in different countries. For example, in Germany, where building codes are very strict, resulting in highly homogenous final roof surfaces, it is common to find single ply PVC membranes being used. In other less developed countries, such waterproofing technologies are not viable and other forms of waterproofing must be tested, proven, and then used. One thing is certain; there must be good communication between the waterproofing industry and the green roof industry to ensure the quality and viability of green roofs. Guarantees must be provided by the water proofer to the client.

Basic Technical Components are Explained in Fig. 1

Growing media choice can affect the roof’s water holding capacity, drainage, and nutrient supply. Nutrient availability is higher for roof garden soils than for extensive roof substrates.

For extensive green roofs, some substrate characteristics have been defined:

Suitable materials are lava aggregates between 2 and 16 mm, pumice, expanded clay, crushed bricks or tiles, basalt gravel, tuff gravel, or gravel.

Growing media criteria have also been described. Good water-holding capacity is defined as more than 45%/volume, pH values should be between 6.0 and 8.5, and the organic matter content should be lower than 90.0 g/l, [1].

Growing media should be mixed locally, in order to achieve regional characteristics and meet

production standards. Each of the different substrate brands have their own trade secrets for finding the best mixture within the material framework described above. Delivering the same quality substrates repeatedly is a must for professional producers of growing media. New mixtures must be tested by certified laboratories. There are testing facilities, where studies analogous to the German testing procedures are conducted, in the USA as well. Growing media for different regions must be engineered to specific local requirements.

Various planting techniques are possible. Seeding is possible, if no danger of erosion exists. Dispersing cuttings and shoots of *Sedum* is a very effective and cost-effective method. For ornamental plantings, mini planters are suitable products. Preproduced turf mats are a well-tested option used in Germany since the 1980s. Similar technologies will be established in the USA and some other countries. Quality criteria for roof plants in Germany follows standards developed by the FLL in Germany and also follow German DIN (Deutschland Industry Norma) standards such as DIN 18916.

If green roofs are to be installed for specific purposes, such as in natural conservation areas where native plants are wanted, more sophisticated standards must be applied.

Suitable vegetation should be selected according to the type and depth of growing media. On shallow layers up to 15 cm in all climates, drought-adapted plant species are the best choice. Succulents are a favorite group in the Northern Hemisphere [36]. Also, in hot humid tropics some succulents, originally from the South African cape region perform well on extensive green roof structures [37]. The reason for this is that these areas may also experience prolonged drought [38]. Xeric gardens with cacti are good choices in the tropics for saving irrigation water. The concept of xeric roof gardens would be choice for dry adapted plant species and an ecological protection layer for the building [39]. The old green roofs in Europe were planted with a type of xeric garden. Environmental education is needed to reeducate the public to understand and appreciate the natural beauty of dry gardens. A public that has come to consider manicured English highly maintained lawn to be the ultimate in visual appeal without realising the obscene damages by the

high input of fertilizer and pesticide to the urban ecosystem causes. Dry green roofs like the Greek oikostegi perform well from the view point of energy savings and rainwater retention as already stated.

The genus *Sedum* is the favorite for many extensive green roof versions. *Sedum* can survive long periods of drought. In temperate climates, *Sedum album* is a species that can exist on nearly no growing media. If the roof has a little shade, *Sedum album* can cover the full roof area. The genus *Sedum* has representatives in various regions of the Northern Hemisphere (Clausen). Furthermore, many cultivars exist and add to the local flora. There are over 600 species of *Sedum* growing naturally in every climate on earth. Consequently, *Sedum* is an ideally suited green roof plant (Figs. 4–7).

Extensive green roofs can be an extension of the surrounding natural ecosystem using endemic local plants. Variation in the depth of the growing media and different types of growing media can be the foundation for a high diversity of plant species and other organisms (Table 3). In order to support plant species which are less competitive, specific maintenance to keep some space on the green roof open is required. Vegetated roofs can also be designed for the specific purposes of some key species of natural flora or fauna [41]. The bird Black redstart is such a key organism for the green roof movement in London. These birds, whose habitat is threatened by urban encroachment, need more habitats to survive and thrive. So-called brown roofs mimic natural environments and provide viable habitat for such birds [42]. Supporting wildlife with specifically designed green roofs is a way to reduce the destructive impact of new developments that encroach on natural habitats.

Green Roofs as Green Corridors in Urban Areas

Green roofs, living walls and back and front yard gardens can become corridors in cities providing pathways of biodiversity. Such green belt systems connect different open urban spaces for plant species and urban animals. The quantity and the quality of these areas should be assessed and modeled in order to enhance future efforts in this direction. In Germany, sites are assigned scores which are called “Biotope area factor” in Berlin or KÖH value in Karlsruhe.



Green Roof Infrastructures in Urban Areas. Figure 4

Sedum album and other Sedum species on the green roof research



Green Roof Infrastructures in Urban Areas. Figure 5

A well-developed multispecies extensive green roof (Neubrandenburg)

In Germany, these scores are the basis for assessing property value and so they are also used to assess measures that can be taken to improve the quality of life in cities.

In Germany, green roofs have become so popular that the percentage of a city that has been greened now can be calculated rather than just talk about the square meter age. Of course, there is still plenty of room for



Green Roof Infrastructures in Urban Areas. Figure 6

Arial view of an ecological neighborhood, all with extensive green roofs, near the lake Constance



Green Roof Infrastructures in Urban Areas. Figure 7

Ecological neighborhood Huckstorf near Rostock with about 50 single and double houses. This is one example of about 180 documented existing ecological settlements in Germany (more: <http://www.oekosiedlungen.de/>)

Green Roof Infrastructures in Urban Areas. Table 3
Types on vegetation and growing media depth

Depth	Type of vegetation	Description
Lower than 5 cm	Mosses, few annual herbs	In most climate regions, this is not enough growing media for a full plant cover
5–15 cm	Typical depth of a green roof growing media	Within this depth, various plant species are suitable, up to 10 cm in most parts of Europe; Sedum and some herbs are the winners. With more than 10 cm depth grass vegetation dominates [40]
15–50 cm	Suitable for turf, herbs, and shrubs	This is the depth suitable for a wide range of ornamental plant species with irrigation. Without irrigation, several herbs and grasses can survive for a long time.
More than 50 cm	Shrubs, small trees	More than 50 cm with irrigation can support shrub and tree vegetation

even more green roofs. The popularization of roof greening in Germany has been achieved in large part due to carrot and stick legislation and incorporation of green roof technology into building codes in much of the country. In Germany, about two third of all building permits are only given when a green roof is included in the application. The German Association of green buildings (www.FBB.de) has conducted three surveys concerning green roof penetration of all German Cities larger than 10,000 inhabitants. Summarized results of these studies are available [43].

Future Direction: A Simulation of Potential Green Roof Benefits to Urban Areas

Each green roof is unique, but new projects can benefit from well-documented existing ones. Detailed documentation, as a case study of a retrofit project, has been prepared for the green roof on the ASLA headquarters in Washington DC [44]. This project includes roof gardens and extensive green roofs. The project is

being monitored to assess the benefits of the installation. This example could be applied around the world. Due to technological advancements, green roofs are now suitable and viable for both retrofit and new buildings [45].

To achieve city-wide benefits, many green roofs must be erected. The potential impact of green roofs on urban environments has been described for Toronto, in a sophisticated study. The aim of the study was to ascertain what the benefits would be derived if all roofs larger than 350 m² were greened in Toronto (about 500 ha). The study found that such a move would result in savings of C\$46,000,000 from storm water amelioration alone. Annually, cost savings of avoiding beach closures were about C\$755,000. An annual reduction in air-conditioning requirements during the summer equaled C\$12,000,000 [22]. This equals an air-conditioning requirement reduction of 2.37 kWh/ m² year.

The benefits of green roofs for New York have been modeled [46] by measurements of the PennState University. This study suggested that significant reductions in heating requirements particularly during the night would result from the installation of green roofs. Also, the study found similar runoff amelioration benefits to Toronto. These two benefits are driving the implementation of green roofs in NYC. Another oft overlooked benefit of the storm water runoff amelioration impact of green roofs is that smaller bore sewage systems can be used meaning substantial economies to local authorities as they rebuild decaying infrastructure. Many other city authorities are developing a healthy interest in green roof research. For instance Seattle, Vancouver, and Washington are all investigating the benefits of green roofs. Rain harvesting potentials of green roofs is another area of interest to local authorities [47, 48].

A combination of green roof technology with other instruments such as living walls using climbers and other technologies opens a wide range of possibilities for greening in Cities [49].

Acknowledgment

Thanks to Olyssa Sterry, MSU, USA, for proof reading the first draft of this paper. Thanks also go to an endless number of friends and colleagues whose contribution to this paper is much appreciated.

Bibliography

Primary Literature

1. FLL (ed) (2008) Richtlinie für die Planung, Ausführung und Pflege von Dachbegrünungen. Bonn, English version: Guidelines for the Planning, Construction and Maintenance of Green Roofing – Green roofing Guideline, 2008 edition
2. Frank A (2009) Planning principle for Sealing Flat Green Roofs – Roof Technology. In: Ansel W (ed) Proceedings of Green Roofs Appl R, Nuertingen, pp 91–98
3. Köhler M, Barth G, Brandwein T, Gast D, Joger HG, Seitz U, Vowinkel K (1993) Fassaden- und Dachbegrünung. Ulmer, Stuttgart
4. Osmundson T (1999) Roof gardens. Norton, New York
5. Rueber E (1860) Das Rasendach. Cotta, München. (reprint 1998) Hannover
6. Ahrendt J (2007) Historische Gründächer – ihr Entwicklungsgang bis zur Erfindung des Eisenbeton. PhD. Technical University Berlin (downloadable)
7. Siqueira VB (2002) Burle Marx paisagens transversas. Cosau Naify, Sao Paulo
8. Nagel W (2009) And compare: the Journal “Häuser”, special issue January “Der Schatz der Moderne - Bauhaus”, www.haeuser.de
9. Nielsen S (2004) Sky gardens. Schiffer, Atglen, PA
10. Peck S (2008) Award winning Green roof design. Schiffer, Atglen, PA
11. Köhler M, Keeley M (2005) Green roof technology and policy development. In: Hoffmann L (ed) Green roofs. Schiffer, Atglen, pp 108–112
12. Minke G (2006) Building with Earth – design and technology for a sustainable architecture. Birkhäuser, Basel
13. Köhler M, Schmidt M (1997) Hof-, Fassaden und Dachbegrünung – Zentraler Bestandteil der Stadtökologie. Landschaftsentwicklung und Umweltforschung 105:1–177
14. FBB, Bohlen R (ed) (2002) Grundsätze zur Pflege und Wartung von Dachbegrünungen. FBB, Saarbrücken. Download www.fbb.de
15. Standard Guide for Selection, Installation, and Maintenance of Plants for Green Roof Systems (2006) Annual Book of ASTM Standards ASTM, Standard E 2400-06. ASTM International, West Conshohocken, PA
16. Standard Practice for Determination of Dead Loads and Live Loads Associated with Green Roof Systems (2005) Annual Book of ASTM Standards ASTM, Standard E 2397-05. ASTM International, West Conshohocken, PA
17. Velazquez R (2008) Green roof calculator. In: Proceedings of the 6th Conference on Green roofs, Baltimore, MD
18. Hämmerle F (2002) Dachbegrünungen rechnen sich. In: Thalacker (ed) Jhrb. Dachbegrünung, pp 18–19
19. Green Values (ed) (2005) Case studies. Green building, growing assets. Royal Institute of Chartered Surveyors (see: www.riscs.org/greenvalue)
20. Hoffman L, Lossvelt G (2007) Viridian green roofs for multifamily affordable housing: reducing upfront costs and creating financing opportunities. 5th Annual Greening roof top conference, Minneapolis, MN, USA (see: www.greenroofs.org)
21. Köhler M, Porsche U (2003) Life cycle costs of Green Roofs – A comparison of Germany, USA, and Brazil. In: Krauter (ed) Rio3.com: Proceedings of world climate & energy event, Rio de Janeiro, Brazil, 1–5 December 2003, pp 461–467
22. Banting D, Doshi H, Li J, Missios P (2005) Report on the environmental benefits and costs of green roof technology for the city of Toronto. Ryerson University, Toronto, p 88
23. Miller T, Liptan T (2005) Update on portlands integrated cost analysis for widespread green implementation. In: Proceedings of the 3rd Conference on Greening Roof tops, Washington, DC
24. LUI STE (2008) Life cycle assessment of green roof systems in Hong Kong, p 96. Dissertation, University of Hong Kong
25. Köhler M (2009) Green roof policies in Germany: successes and Outlooks. In: Proceedings on Cities alive, Toronto, Oct 2009
26. Tan PJ (2008) Environmental, economics and social benefits in adopting skyrise greenery. In: Proceedings of Regional Seminar CUGE, 22 Oct 2009
27. Townshend D, Duggy A (2007) Study on green roof applications in Hong Kong. Urbis Limited, Hong Kong, pp 1–153
28. Lee UH (2009) Green roof development in Seoul, Korea. In: Proceedings of Cities alive, Toronto, Oct 2009
29. Koshimizu H (2009) Green roofs in Japan. In: Proceedings of Cities alive, Oct 2009 (in press)
30. Chinadaily, posted: 2008-08-12; Shanghai closes in on green roof target. Shanghai Post http://www.chinadaily.com.cn/china/2008-08/12/content_6926333.htm (site visited July10s, 2009)
31. Hämmerle F (2002) Der Markt für grüne Dächer wächst weiter. In: Thalacker (ed) Jhrb. Dachbegrünung, pp 11–13
32. Koshimizu H, Lee H (2007) The regulation regarding the rooftop greening in the East Asian Cities. In: Proceedings of 5th annual conference on Greening rooftops for sustainable communities, Minneapolis, MN
33. Köhler M, Malorny W (2009) Wärmeschutz durch extensive Gründächer. In: Venzmer H (ed) Europäischer Sanierungskalender 2009, pp 195–212. Beuth, Berlin
34. Maurer E (2009) Successful green roof policies in Linz since 1985. In: IGRA Proceedings, pp 169–172
35. Keeley M (2007) Transatlantic exchange and sustainable urban development: transferring stormwater policies and technologies from Europe to the United States, p 259. PhD, Technical University of Berlin
36. Snodgrass EC, Snodgrass LL (2006) Green roof plants. Timber, Portland
37. Tan PJ, Sia A (2005) A selection of plants for green roofs in Singapore. CUGE, Singapore
38. Tan PY (2009) Understanding the performance of plants on non irrigated Green Roofs in the Tropics using a Biomass yield approach. Nature in Singapore. (<http://rmbr.nus.edu.sg/nis>)

39. Köhler M (2009) How Green should a green roof be. 7th Greening Rooftops for sustainable Communities. Atlanta, 5–7 June 2009
40. Köhler M (2006) Long term vegetation research on two extensive roofs in Berlin. *Urban Habitat Urban Habitats* 4(1):3–26
41. Brenneisen S (2006) Space for urban wildlife: designing green roofs as habitats in Switzerland. *Urban Habitats* 4:27–36
42. Gedge D, Newton J, Cradick K, Cooper P, Partington T, Bramfield T, Kendall J (2008) Living roofs and Walls. Technical Report: Supporting London Plan Policy. London 575
43. Köhler M (2010) Green Roof Policy in German: successes and Outlooks. In: Proceedings on Cities alive congress, Urban policy (in press)
44. Werthmann C (2007) Green roof – a case study. Princeton, New York
45. Weiler SK, Scholz-Barth K (2009) Green roof systems. Wiley, Hoboken
46. Rosenzweig C, Gaffin S, Parshall L (ed) (2006) Green roofs in the New York Metropolitan Region. Research Report Columbia University. <http://www.ccsr.columbia.edu/cig/greenroofs/>
47. Dunnet N, Clayden A (2007) Rain gardens. Timber press, Portland
48. Dreiseitl H, Grau D (2006) Wasserlandschaften. Birkhäuser, Basel
49. Köhler M (2008) Green facades – a view back and some visions. *Urban Ecosyst* 11:423–436

Books and Reviews

- Dunnet N, Kingsbury N (2008) Planting green roofs and living walls, 2nd edn. Timber, Portland
- Ernst W (2005) Dachabdichtung Dachbegrünung. IRB-Fraunhofer Gesellschaft, Stuttgart
- Köhler M, Barth G, Brandwein T, Gast D, Joger HG, Seitz U, Vowinkel K (1993) Fassaden- und Dachbegrünung. Ulmer, Stuttgart
- Krupka B (1992) Dachbegrünung. Ulmer, Stuttgart
- Snodgrass EC, Snodgrass LL (2006) Green roof plants. Timber, Portland
- Weiler SK, Scholz-Barth K (2009) Green roof systems. Wiley, Hoboken
- Jodidio (2009) Green Architecture now, p 416. Taschen, Hongkong

Recommended and Selected Internet Links

www.worldgreenroof.org
www.greenroofs.org
www.fl.de
www.greenroofs.com
http://www.epa.gov/npdes/pubs/gi_supportstatement.pdf
http://www.usmayors.org/urbanwater/policyres_06c.asp
<http://www.ecos.org/content/policy/detail/2861/>
http://cfpub.epa.gov/npdes/home.cfm?program_id=298
<http://greenvalues.cnt.org/calculator>
<http://freitext.ub.tu-berlin.de/gartenkunst.html> (garden historic documentation on older Green roofs in Journals of the 1920th)
<http://www.oekosiedlungen.de/>
<http://www.oikosteges.gr>

Green Roof Planning in Urban Areas

STEPHAN BRENNISEN¹, DUSTY GEDGE²

¹Life Sciences und Facility Management, Wädenswil, Switzerland

²Livingroofs.org Ltd, London, UK

Article Outline

Definition of the Subject and Its Importance

Introduction

The Headline Benefits of Green Roofs to Urban Planning and Climate Change Adaptation

Green Roofs and Pollution Removal from Stormwater Runoff

Role in Urban Planning

Green Roofs in the Rest of the World: A summary

Bibliography

Definition of the Subject and Its Importance

Green roofs are vegetated substrate layers on top of the waterproof membrane of the conventional roof surfaces of buildings. Once the concern that plants could negatively affect plants was overcome, by the use of improved membrane layers that are protected from root penetration, green roofs became more popular, starting in the late 1970s in Germany, Austria, and Switzerland.

As the twenty-first century has been deemed by the UN as the century of the megacity with predictions that nearly 75% (?) of the population of the world will live in cities by the end of the century, urban areas will increasingly need to be planned and developed to take into consideration the ecological and environmental health of the urban climate to ensure cities can cater for people. This is also a pressing matter in as it is recognized that there is a need to adapt cities to the negative effects of climate change. These issues have been central to the work of professionals in the field of green roofs over the last 20 years and increasingly the technologies and approaches of green roofs have been shown to have a positive effect on a number of issues facing cities and mega-cities in terms of ecological and environmental health and in helping cities adapt to climate change, reduction of the urban heat island effect, reduction in localized flash floods, air and

noise pollution and increasing biodiversity in the urban realm. These processes are part of an evolving approach to planning referred variously as green infrastructure or ecosystem services. Urban greening has therefore become an increasingly important approach for planners and is likely to be the one that will predominate in the future.

The Swiss architect Le Corbusier set the installation of green roofs in the form of roof gardens, as one of his five principal requirements in establishing a new architecture in the early twentieth century (Busse 2000). Although his work became well known, it took a long time after Le Corbusier's first steps to the implementation of green roofs as an ecological measure to be absorbed into mainstream urban planning. Green roofs became more popular as a part of ecological construction in the 1970s, and in the early 1980s, a number of cities implemented nascent strategic planning approaches to encourage the uptake of green roofs. Elsewhere a series of pilot projects were initiated to explore the ideas and approaches to green roofs in urban areas. The main drivers for the implementation of green roofs in the 1970s' and 1980s' research also suggested that green roof provided a broad range of environmental benefits, such as energy savings (less winter heating), reduction in storm water runoff, and overheating, but there was also a recognition that green roofs also promoted health and well-being and were an essentially element in an "ecological" approach to construction.

Introduction

Green roofs are vegetated layers that sit on top of the conventional roof surfaces of a building. Usually a distinction is made between "extensive" and "intensive." These terms refer to the degree of maintenance the roofs require. Intensive green roofs are composed of relatively deep substrates and can therefore support a wide range of plant types: trees and shrubs as well as perennials, grasses, and annuals. As a result, they are generally heavy and require specific support from the building. Intensive green roofs (what most people think of as roof gardens) have in the past been rather traditional in their design, simply reproducing what tends to be found on the ground, with lawns, flower beds, and water features. However, more contemporary

intensive green roofs can be visually and environmentally exciting, integrating water management systems that process waste water from the building as well as storing surplus rainwater in constructed wetlands. Because of their larger plant material and horticultural diversity, intensive green roofs can require substantial input of maintenance resources – the usual pruning, clipping, watering, and weeding as well as irrigation and fertilization.

Conversely, the green roofs that have received the greatest interest recently are extensive green roofs. They are composed of lightweight layers of free-draining material that support low-growing, tough drought-resistant vegetation. Generally the depth of growing medium is from a few centimeters up to a maximum of around 10 cm. These roof types have great potential for wide application because, being lightweight, they require little or no additional structural support from the building, and because the vegetation is adapted to the extreme roof top environment (high winds, hot sun, drought, and winter cold), they require little in the way of maintenance and resource inputs. Extensive green roofs can be designed into new buildings, or "retrofitted" onto existing buildings.

The Headline Benefits of Green Roofs to Urban Planning and Climate Change Adaptation

Green roofs provide a range of benefits to urban areas. However, there is an increasing interest in three particular headline benefits, namely, amelioration of the urban heat island effect [UHIE]/thermal performance of buildings, reduction in stormwater flows, and providing habitat for wildlife in cities.

The first two are particularly relevant in terms of climate change as there is a recognition especially in temperate and Mediterranean climates that climate change is likely to increase summer temperatures in cities, thus increasing the negative effects of the UHIE (and potential increase in death caused by heat excess) and an increase in the need for air-conditioning (and therefore increase in energy/carbon use). The other prediction is that climate change will lead to increase in intense summer storms leading to flooding in urban areas with a negative effect on both the ecological and economic conditions within urban areas. In many cities, these predictions are actually happening already. Therefore, there

is a pressing need to adapt urban areas to ameliorate both effects and thus ensure the ongoing ecological health and economic well-being of urban areas.

Urban Heat Island and Climate Change Adaption

The benefits green roofs can bring in terms of climate change adaptation can be significant, especially in terms of the urban heat island effect and the energy balance of a given building.

Conventional roofing surfaces, including recreational roofs, absorb sunlight and heat up quickly. The absorption of radiation and the release of the radiation back to the atmosphere during the night is a major factor in UHIE. Where a building has poor insulation and poor ventilation, this can lead to increased use of air-conditioning and therefore increased energy use [1].

Building Energy Balance

The use of vegetation on the roof surface ameliorates the negative effects of conventional roofing surfaces by absorbing heat and using this heat in evapotranspiration. The process of evapotranspiration is an important element in reducing UHIE and provides benefit to the individual building by cooling/insulating the spaces beneath the roof.

Studies have shown that the membrane temperature beneath a green roof can be significantly lower than where the membrane is exposed. Table 1 shows the average temperatures under the membrane of a conventional roof and that of membrane under green roofs in a study undertaken at Nottingham Trent University.

Green Roof Planning in Urban Areas. Table 1 Study of temperatures under membranes of a conventional and a green roof

	Winter	Summer
Mean temperature	0°C	18.4°C
Temperature under membrane of conventional roof	0.2°C	32°C
Temperature under membrane of green roof	4.7°C	17.1°C

www.greenroofs.co.uk

Another study in Ottawa, Canada, by the National Research Council of Canada noted that temperature fluctuations during spring and summer on a conventional green roof were of the order of 45°C while under a green roof the fluctuations were in the order of 6°C [6].

The positive effect on the temperature of the membrane under a green roof not only protects the membrane from the effects of UV, frost, and sunlight, but also moderates the heat flow through a building by shading, insulation, evapotranspiration, and thermal mass.

Urban Heat Island Effect and Indirect Energy Savings

- “Summers by 2050 will be 1.5–3.5°C hotter. . .in central London the urban heat island currently adds 5–6°C to summer night time temperatures and will intensify in the future.” *London’s Warming. The Impacts of Climate Change on London* www.london.gov.uk

Urban areas have a higher average temperature than surrounding rural areas; this difference in temperatures is called the urban heat-island effect. An increase in the UHIE is likely to lead to increased air pollution and increased use of air-conditioning.

As has been noted already that green roofs do not store heat like conventional roof surfaces but dissipate heat through evapotranspiration. Such a process can have a significant effect on helping to reduce the impact of the UHIE.

A modeling scenario undertaken in New York by the New York Heat Island Initiative determined that providing 50% green roof cover within the metropolitan area would lead to an average 0.1–0.8°C reduction in surface temperatures. It is noted that for every degree reduction in the UHIE roughly 495 million Kilo Watt hour of energy would be saved. The same study also looked at various mitigation strategies other than green roofs, including urban forestry and cool roofs and noted that “living roofs” provided greater benefits than white or “cool” roofs. It was clear from the study that a combination of various mitigation strategies for UHIE including green roofs should be considered by the city. <http://ccsr.columbia.edu/cig/greenroofs/index.html> www.nyscrda.org/programs/environment/emep/project/6681_25/6681_25_project_update.pdf

A study in Toronto estimated that there are 50 million square meters of potential roof space in the city of Toronto that could be greened. Overall it was estimated that the effect of greening the city's roof tops would lead to 0.5–2°C decrease in the UHIE.

A reduction of this magnitude would, the study estimated, lead to indirect energy savings citywide from reduced energy for cooling of \$12 million, equivalent to 2.37 kWh/m² per year [0.001 CO₂ emissions t/m²]. <http://www.toronto.ca/greenroofs/findings.htm>

Energy Savings

Although green roofs do provide potential energy savings by providing building insulation these are often considered difficult to assess due to the varying climatic conditions throughout the winter months, and will be minimal on already well-insulated buildings. However, during summer months, green roofs can have a significant effect on spaces beneath them in terms of cooling.

Studies in Germany have provided various estimates. Figures provided by Zinco estimate that 2 L of fuel oil are saved per square meters per year. A more recent published study on domestic buildings with flat roofs suggests that there is a 3–10% winter saving on fuel bills. The results of the study suggest that there is a maximum saving of 6.8 kWh/m² [calculated CO₂ emission tones saving of 1.5 kg/m²] and a minimum saving of 2.0 kWh/m² [calculated CO₂ emission tones saving of 0.44 kg/m²]. This study did not consider at any summer savings due to cooling [5].

The Toronto study, already referred to above, estimated that the direct energy savings citywide as a consequence of whole scale greening through reduced energy for cooling would be in the order \$22 million, equivalent to 4.15 kWh/m² per year [0.001 CO₂ emissions t/m²]. The study also concluded that there would be a reduction in peak demand in the order of 114.6 MW, leading to fossil fuel reductions in the region of 56,300 metric tons per year.

The only information for a building in London suggests that an 850 m² retrofitted green roof onto paving in Canary Wharf has seen an estimated reduction of 25,920 kWh in a year, through a reduction in heating and cooling of the spaces below the roof [*per com Livingroofs.org*].

Green Roofs and Photovoltaic Solar Panels

The combination of green roofs and PV Solar Panels provide multiple benefits. Photovoltaic panels at roof level are known to work more efficiently when installed on a green roof rather than a conventional surface. The green roof element not only saves energy during the summer time [see above] but can increase efficiency of PV by reducing fluctuation of temperatures at roof level and by maintaining a more efficient microclimate around the PV panels. The performance of photovoltaic panels is lowered by 0.5% per °C above or below 25°C. The green roof is better at maintaining the ambient temperature of 25°C [3].

By reducing the temperatures around the PV and by helping reduce the need for air-conditioning in spaces beneath the green roof, the combination of the technologies should be as one of positive interaction and not one of competition in terms of use of roof space [4].

Benefit of Green Roofs to Stormwater Runoff, Flash Floods, and Pollution Removal

Background The combined impact of ongoing development within urban areas and climate change has created higher peak stormwater flows, leading to an increased occurrence of downstream flooding and pollution. Intensive summer storms are becoming, and are likely to become, more prevalent. Such rainfall events, especially in the summertime in temperate zones, can cause the current stormwater system to become overburdened causing localized flooding. The consequence of these events can have both a negative economic and ecological effect. For example a “freak” summer storm in 2005 in West London deposited such a large amount of rainwater that caused raw sewage to be released into the River Thames. This pollution caused large-scale ecological damage to the river's ecosystem, including significant fish mortality. Another series of summer storms in July 2009 in London led to closure of parts of the transport system within the capital causing offices, shops, and businesses to be closed and a reduction in trade/business.

In many parts of the world, sustainable drainage systems (SUDS) are now required to minimize the impact of both new and existing development.

They are designed to both manage the adverse environmental consequences resulting from urban stormwater runoff and also contribute to environmental enhancement wherever possible. The use of green roofs can provide a pivotal role in achieving this as they successfully achieve source control, which is the fundamental concept of SUDS, i.e., the control of rainfall at or as close as possible to its source.

Positive Effect of Green Roofs on Runoff Rates and Volumes of Stormwater

Around 30–40% of rainfall events are sufficiently small that there is no measurable runoff taking place from greenfield areas (it all infiltrates or evaporates). In contrast, runoff from developed areas takes place for virtually every rainfall event. This means that streams and rivers are more subject to overload. In addition, whereas for greenfield areas, small events would be treated through natural filtration processes, development runoff can flush surface pollutants directly into the receiving waters. Where it is possible to provide replication of this natural behavior (described as interception storage) and to prevent runoff of up to 5 mm, this should be provided.

By using green roofs as a source control technique, the volume of surface or underground attenuation storage can be reduced considerably. This can be particularly important in dense urban developments where space for surface level SUDS components such as ponds will be limited. It is also an important consideration when looking at the true cost implications of installing a green roof as the reduction in underground drainage infrastructure must be taken into account as well as the reduced number of downpipes and smaller pipe network, etc.

When rain falls on a green roof, it will first pass into the substrate and possibly pass through until the adsorbancy of the soil is activated (although through-flow will generally be low). It is then absorbed by the substrate (and possibly the drainage layer) and taken up by plants in the same manner as a greenfield site.

For most small storms the rainfall is removed by evapotranspiration. Only when the soil is fully saturated will water percolate through to the underlying drainage layer in significant volumes.

The processes involved in the operation of a green roof are (Tarr 2002):

- Retention of rainwater in substrate and drainage layers
- Uptake of water and release by plants as vapor (transpiration)
- Uptake of water and biochemical incorporation by plants (photosynthesis)
- Evaporation from substrate due to wind and sun

There is a wealth of published information that demonstrates the performance of green roofs in attenuating stormwater runoff by reducing peak flow rates and volumes.

Although there is a variation in performance depending on rainfall patterns, this is no different to other SUDS components such as pervious pavements, or indeed greenfield catchments.

The benefits of a green roof in terms of drainage can be summarized as follows:

1. A green roof will typically intercept the first 5 mm and more of rainfall (provide interception storage).
2. The amount of stormwater stored and evaporated is primarily dependent upon the depth of the growing medium and type of planting. In the summer, a green roof can typically retain between 70% and 80% of the runoff (Livingroofs.org 2004).
3. It has been demonstrated that in Germany between 40% and 100% of rainfall can be retained – depending upon the season (Tarr 2002).
4. Seventy-five percent of rain falling on extensive green roofs can be retained in the short term and up to 20% can be retained for up to 2 months (English Nature 2003).
5. As the rainfall events become longer or more intense, the positive effect of a green roof remains as there is still a significant reduction in peak runoff rates. This increase in the “time of concentration” means that a green roof will be beneficial throughout a wide range of rainfall conditions [2].
6. The above benefits collectively mean that by incorporating a green roof into new development, there will be a reduction in the amount and cost of the overall drainage infrastructure required to serve that development.

Green Roofs and Pollution Removal from Stormwater Runoff

Green roofs retain, bind, and treat contaminants which are introduced to the surface either as dust or suspended/dissolved in rainwater.

The London Ecology Unit (1993) stated that 95% of heavy metals are removed from runoff by green roofs and nitrogen levels can also be reduced. In addition, Auckland Regional Council (2003) advise that green roofs are accepted as removing 75% of total suspended solids.

The roofs showed a reduction in nitrogen (total discharge from green roofs of between 10 and 80 mg/m³ during the monitoring period) and phosphate was also removed from the runoff (total discharge of between 75 and 100 mg/m³). The total discharges of nitrogen and phosphate from the conventional roof were 265 and 145 mg/m³, respectively.

Role in Urban Planning

Examples from

London During the late 1990s, a group of nature conservationists and ecologists were actively involved in a major regeneration program in South-East London. The Creekside Regeneration Programme was a government-funded program to rejuvenate an economically deprived area of Deptford in the London Boroughs of Lewisham and Greenwich. Part of the preparatory works included an ecological assessment of the value of the area for wildlife. Although a rundown postindustrial landscape the group discovered that, although not the most visually appealing landscape there were a number of interesting species, including one – the black redstart (*Phoenicurus ochruros*), which were of significant ecological value.

Although the use of green roof technology had been tentatively used in the 1970s and a number of innovators had tried to raise the profile of the technology through example projects and a book published by the London Ecology Unit [Building Green], the technology had remained relatively marginal.

However, the work of the group in Deptford pushed developers and planners to consider the use of green roofs as mitigation for wildlife habitat, especially for

the aforementioned bird. However, with little information on the value or the technical aspects of green roofs, the process was a hard task. Convincing the relevant professionals, architects, consultants, and planners that this was both technically and an ecologically sound method of roofing was an uphill struggle. In effect, the group was on a process of change, changing people's habitats and assumptions and professional attitudes to vegetation on buildings. Fortunately, in early 2000, the author contacted the Swiss Landscape Department, which led to the first contact with Dr. Stephan Brenneisen and his innovative work on green roofs and biodiversity.

Thus was born the UK+CH green roof partnership. Although the Swiss regulatory process and attitude to green roofs was on a different level at the time, being one of proactive engagement, the Swiss experience has been a major influence on bringing green roofs into the mainstream in London and elsewhere in the UK. The relationship has been one of the major influences in transforming green roofs from a fringe sustainable technology to one that is now incorporated in both strategic policy in London, both at the regional and the local level. Although the agenda for green roofs as moved forward from an initial interest in biodiversity and nature conservation issues, to other themes, such as thermal performance, urban heat island amelioration, and sustainable urban drainage, at its heart the provision of quality green roof habitat for rare invertebrates and birds still remains, and is still a major driver in terms of implementation.

The Black Redstart Although relatively common on the continent and especially in Switzerland, this bird is at the northwesterly edge of its range in the UK. As it is a relatively recent addition to the UK bird fauna, it is vulnerable to competition from more established species, such as the Robin (*Erithacus rubecula*). With less than 100 pairs annually thought to breed in the UK, it is therefore protected under national wildlife legislation. It is relatively unique for a UK protected species as it is found in the main on industrial wastelands and in an urban context. It needs quite sparsely vegetated landscapes, similar to alpine scree slopes and plateaus.

In 1997, a group of ecologists found a number of pairs breeding along Deptford Creek. As the three pairs that were found in the area constitute more than 1% of the national population, there was a degree of pressure, through planning regulation, to ensure development did not have a negative impact on the species. Wherever possible there is an obligation to “enhance or mitigate” development plans for protected species. Thus was born a move to ensure that new developments catered for the black redstart. One of the first projects to be impacted on was the Laban Dance Centre, designed Herzog de Meuron.

The “rubble” roof was the first of many green roofs installed for black redstarts in London. Thirteen years ago this was a real milestone. But now, after much work, green roofs for biodiversity have become embedded.

However, the nature conservationists involved in promoting green roofs had very little knowledge or understanding of the know-how and the technical arguments to pressure developers to deliver habitat at roof level. Most of their thinking was intuitive. Furthermore many professional ecologists viewed the “idea” as unfounded and lacking in any hard and fast data to back up the intuitive reasoning of group. The Swiss connection would change that providing a greater technical understanding of green roofs supported by detailed research.

The Ecological Context While the black redstart started was the starting point for a renaissance in interest in green roofs in London and increasingly across the rest of the UK, a further issue raised its head with the publication of a strategic white paper that set out the vision for development in the UK but which nature conservationists saw as having potentially negative impacts on a swathe of habitats, that, although postindustrial, had become important refuges for a whole swathe of rare and endangered species. In 2000, the UK Government published the Urban White Paper. In the context of this story, the white paper targeted the use of brownfield and postindustrial land as the primary source of land for new development. In London and also in the Thames Gateway Development region, large areas of old factories, dockyards, disused refineries, power stations, and other land become the prime focus for economic regeneration. However, at

the same time, these rather ugly and untidy landscapes were increasingly being recognized by urban ecologists as some of the most important sites in the UK for rare invertebrates. Much of this fauna had once been prevalent in well-drained and low-nutrient farmland that had since been improved. In general many postindustrial sites reflect this habitat characteristic. Thus, over time, much of the fauna and flora had colonized these as they were left to nature. In fact one such site in South Essex, Canvey Wick, has been celebrated as “the Amazonian rainforest for rare invertebrates in the UK.” Thus was born a conflict between the needs for economic regeneration and an ecology once overlooked but of national importance. At the time of writing a number of other similar sites have become the focus of concern from conservationists, notably the Isle of Grain lorry park (<http://www.guardian.co.uk/environment/2011/apr/17/isle-of-grain-wildlife-paradise>).

Climate Change: A Shift in Emphasis During the mid-2000s, there was a distinct policy shift as the climate change agenda became more prevalent. Where the London Mayor had previously been concerned with “fuel” poverty a series of extreme hot summers shifted the emphasis to also include concerns regarding “cool” poverty and the likely impacts of increased summer temperatures and the negative effects of the Urban Heat Island. The London Climate Change Partnership stated in 2002 that “Summers by 2050 will be 1.5–3.5°C hotter...in central London the urban heat island currently adds 5–6°C to summer night time temperatures and will intensify in the future” [6].

Thus, green roofs were seen to be a potentially crosscutting technology and multi-beneficial solution that could help the capital meet the challenges of climate change and help the city adapt to the likelihood of higher summer temperatures and increases in intense summer storms [leading to localized flash floods]. These fears have been realized to a certain extent with the exceptionally hot summers of 2005 and 2006 and extensive flash flooding of 2007.

However, London has both a central strategic authority [the Greater London Authority] and also 32 local boroughs [LB] with responsibility for their area. A number of these boroughs, notably city of London,

Islington, Lewisham and Tower Hamlets, were actively promoting and ensuring that green roofs were installed for nature conservation.

The New London Plan In 2007, the author, along with colleagues, was commissioned to write a technical review of green roofs and green walls. This report was to support the change in policy on green roofs from one of “encouragement” to “expectation.” The report reviewed all the technical data available for a range of benefits including reduction in the urban heat island, thermal performance, storm water attenuation, biodiversity, and amenity. It also reviewed city policies elsewhere in the world. The technical report, published in late 2007, led to a distinct policy on living roofs and walls in the revised London Plan published in March 2008.

The new London Plan Living roofs and wall [Policy statement 4a 11] states:

The Mayor will and the boroughs should expect all major developments to incorporate living roofs and walls, where feasible and reflect this in Local Development Framework policies. It is expected that this will include roof and wall planting that delivers as many of these objectives as possible:

- Accessible roof space
- Adapting and mitigating for climate change
- Sustainable urban drainage
- Enhancing biodiversity
- Improved appearance

Boroughs should also encourage the use of living roofs in smaller developments and extensions where the opportunity arises [5].

Although it is too early to see what the effect of the new policy, green roofs have been delivered on an increasing scale in the capital over the last 8 years and the new Plan should lead to an increase in delivery.

A London Audit Livingroofs.org undertook an audit of green roofs in London in 2004. The total area of green roofs was no doubt underestimated due to the challenges of accessing information from architects, developers, and companies. The audit estimated that 76,682 m² of green roofs had been installed in the Greater London area. Some of these roofs dated back to 1932! In 2008, a further audit was undertaken to

assess the amount of green roofs installed between January 2004 and December 2008. This used data provided by a number of green roof companies active in the UK and is estimated that the figures provided represent 80% of the roofs installed during that period. Over 420,00 m² were installed by the companies suggesting that near to 500,000 m² were actually installed – equivalent to twice the size of Hyde Park and Kensington Gardens combined.

It is important to note that the majority of these green roofs were installed in the central core of the city. Furthermore as the table below highlights that the area with the largest percentage of installed green roofs was the London Borough of Islington. This borough had actively promoted the use of green roofs through its planning department, thus demonstrating that urban planning departments that embrace and promote these activities can have immediate effect in the implementation of green roofs.

This area of green roofs installed between 2004 and 2008 was prior to the publication of the New London Plan, and the positive effect of this strategic planning policy on green roof implementation is yet to be assessed but is likely to augment the delivery of green roofs on new developments in the Capital.

A Green Roof Toolkit Although London now has a distinct policy on green roofs and a large number of green roofs have been installed, there has been limited guidance on how green roofs should be installed in London.

The Environment Agency, a key statutory consultee on new developments, is responsible for flood defense and rivers. Its remit covers flooding, climate change, and biodiversity. In 2008, the Thames region of the Environment Agency commissioned the author to provide a detailed toolkit to guide developers to what the Agency expected green roofs to provide and how. Importantly this work also managed to persuade engineers within the Agency to accept that green roofs do provide significant benefits as a source control mechanism in the sustainable urban drainage train [4].

Greening Existing Building Stock With the new London Policy and activities by the 32 LBs, the argument for green roofs on new developments looks

positive in the London area. However, one of the key concerns for the London Climate Change Partnership [LCCP] is the need to adapt existing buildings to the challenges of climate change. Reduction of the Urban Heat Island is an important element and, as has been recognized elsewhere, there is a need for a 10% increase in urban green spaces in UK cities to combat climate change. In London, especially in the central activity zone [CAZ] where space is limited and the effects of the Urban Heat Island are most pronounced, roofs will be very important in delivering green space.

The Need to Retrofit The implementation of green roofs in new developments within London is now a mainstream ecological construction method as has been outlined above. However, it is also recognized that there is a need to retrofit green roofs on the existing building stock, especially in CAZ. The report published in 2008 by the London Government estimates that 325 of the land area of Central London consists of roofs that could be retrofit with green roofs in the future without additional structural work to facilitate such an approach. Therefore, the area of roofs within 6 km circle centered on Trafalgar Square that could be greened would constitute a total area of 10million square meters.

The burning issue for planners is how to achieve wholesale retrofitting of green roofs on existing buildings. The London Government commissioned a report, economic incentive schemes, for retrofitting London's existing homes for climate change impacts, which outlines how this might be achieved in terms of incentives and the monies that would be needed to achieve. The report also provided detailed figures on how much such a program would cost and the environmental benefits that it would provide:

- “A scheme for four inner city areas – Cannon Street, Oxford Street, Tottenham Court Road, and Canary Wharf – with a green roof area of 226,750 m² would cost around £4 million and provide environmental benefits 17 worth £4 million.
- A wider scheme covering the city of London, part of the London Borough of Hackney, part of the London Borough of Tower Hamlets, and part of the
- A scheme in the West End with a green roof area of 3.2 million square meters would cost around £55.5 million and provide environmental benefits worth £55.5 million” (Table 2).

Green Roof Planning in Urban Areas. Table 2

Area	Potential roof area that could be greened (m ²)	NPV of environmental benefits	Size of scheme
Four inner city areas – Cannon Street, Oxford Street, Tottenham Court Road, and Canary Wharf	226,570	£4 million	£4 million
Four larger sample areas – city of London, part of the London Borough of Hackney, part of the London Borough of Tower Hamlets, and part of the West End	3.2 million	£55.5 million	£55.5 million

Economic incentive schemes for retrofitting London's existing homes for climate change impacts www.london.gov.uk/lccp/publications/docs/lccp-eco-incentives.pdf

In a further development since the publication of the report, the idea of retrofitting green infrastructure and in particular green roofs in London is being taken forward by Business Improvement Districts [BID], in particular the Victoria [VBID] area in the South West of the Central Activity Zone. VBID commissioned a report and survey of the inner and outer areas that constitute the BID to assess where and what kinds of green infrastructure could be retrofitted. As has been mentioned early, the driving force for the commission was the effects of climate change. The VBID area was one of the key areas affected by flashfloods in the first week of July 2009 with the closure of Victoria Mainline station and VBID recognizes that this led to a significant loss of business within the zone.

The report highlights the issues of flood management within the BID area:

- One of the key environmental challenges in the Victoria BID is the need to reduce instances of flooding at Victoria Station during periods of heavy rain. These instances are likely to become more frequent, and the UK Climate Impacts program predicts that the average winter rainfall could increase by between 12% and 16% by 2050 and by 16–26% by 2080 (UK Climate Impact Projections website, accessed August 2010: <http://ukclimate-projections.defra.gov.uk/content/view/2200/499/>). Well-designed GI can alleviate the risk of flooding by retaining water.
- Environment Agency data indicates that extensive areas of Westminster are prone to fluvial and tidal flooding, including many parts of Victoria, as shown in Fig. 1. This includes the area directly outside the entrance to Victoria station, and either side of Bressenden Place. Much of the Victoria BID is identified as a Critical Surface Water Flood Location in the recent Westminster Strategic Flood Risk Assessment (SFRA). The SFRA predicts that this surface water flooding will be exacerbated by the predicted effects of climate change, and will affect an even greater area of the Victoria BID. The Westminster SFRA also indicates that a breach of the Thames Barrier would result in flooding which would extend as far as the eastern part of the Victoria BID.

An audit of the roofs was also under and each roof assessed as to whether it could potentially be greened and what type of green roof in terms of depth could be achieved.

The flat roof audit identified approximately 29 hectares of roof area within both the outer and inner cores of VBID area 4.17. Of these 29 hectares of roof, over 25 hectares had the potential to support a living green roof habitat. Of this 25 hectares, 18% had the potential to support a green roof in keeping with guidelines of biodiversity as outlined in the Environment Agency Green roof toolkit, 55% moderate potential and 42% low potential.

Since the publication of the report, other similar audits are being considered elsewhere in London and also beyond the capital in UK.

Both the London Government's report and VBID report demonstrate the commitment in theory to retrofitting green roofs. However the major task will be to ensure that the theory is put into practice. There are now many examples of green roofs that have been retrofitted in London, most of these have been done through either funding streams associated with biodiversity or where commercial companies have engaged with Livingroofs.org to specifically deliver seminal projects at their own expense, such as the highest green roof in the UK on top of Barclays Headquarters in Canary Wharf, London. Over the last 4 years a project funded by the SITA NATURE ENHANCED has been ongoing in London. This has led to just under 2,000 m² of green roofs being installed on existing buildings throughout the capital for rare invertebrates. One of these projects, 55 Broadway, recently won the Sustain Award for Biodiversity.

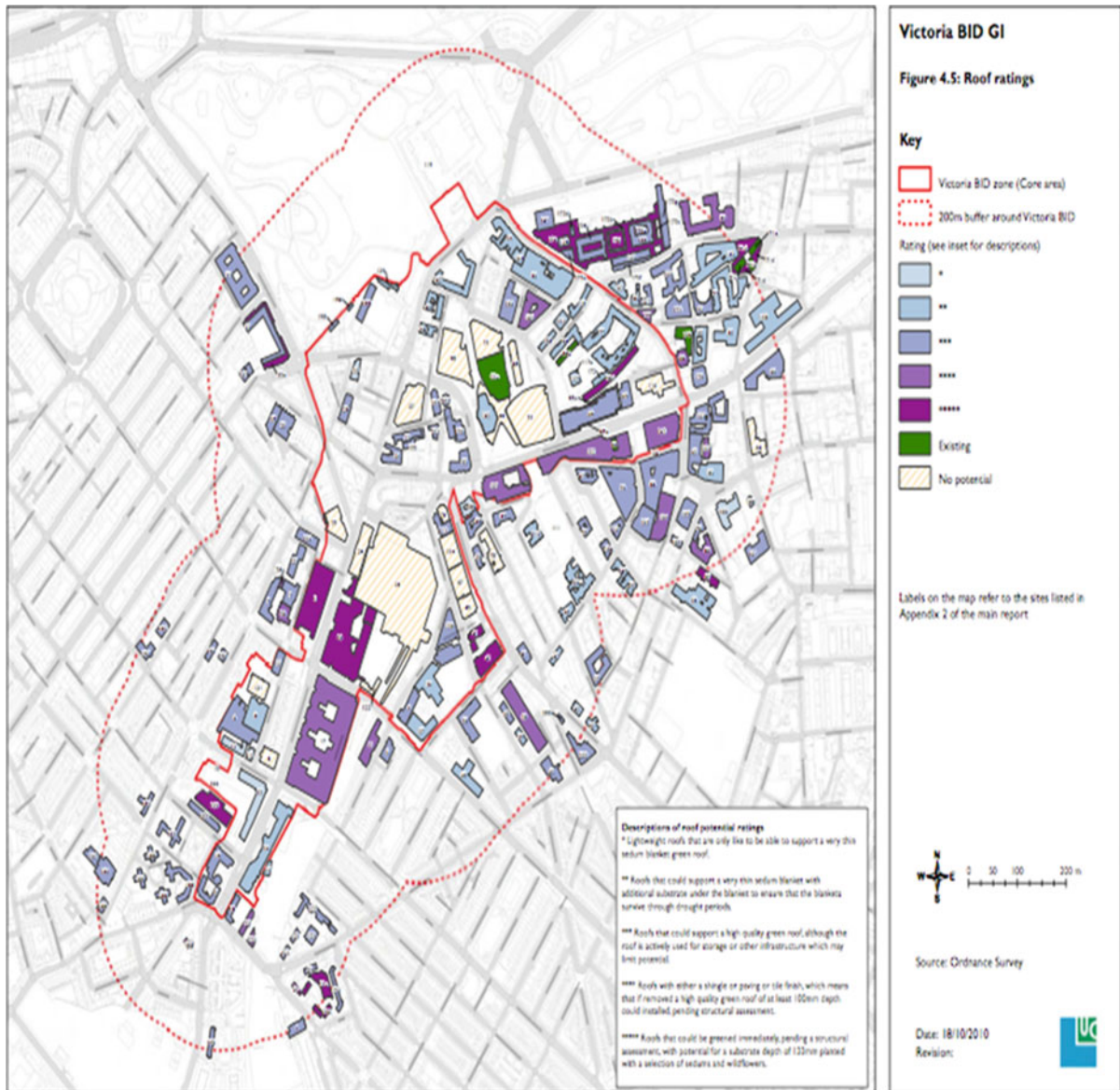
A further development that is underway in London is the retrofitting of a number of green roofs as part of the Drain London program (<http://www.london.gov.uk/who-runs-london/greater-london-authority/directors-decisions/dd421>). The first two will be installed in the summer of 2011. Part of the money allocated for this project is utilized to have the roofs monitored for their positive impact on intense summer storms.

One of the roofs to be funded

The Importance of Policy and refined Planner Tools for green roofs

There is still unfortunately a tendency on designers, whether they be architects or landscape designers, to seek out mechanistic product-based solutions. While most of the

The Future in the UK The development of a unique London green roof policy has been driven by a number of specific factors. Urban ecologists and their issues certainly provided the initial impetus for policy activities and innovations. The climate change agenda and how a large city like London adapts has certainly galvanized the need for a policy. What is needed in the future is a greater understanding and improved planning tools to ensure that good green roofs are installed to meet the cross cutting benefits that a city like London requires. It would be hoped that the Environment Agency Toolkit sets a precedent that at regional level [GLA] and a local [LBs] green roof



Green Roof Planning in Urban Areas. Figure 1

policies and conditions can be refined to ensure the quality of green roof implementation. There is still unfortunately a tendency on designers, whether they be architects or landscape designers, to seek out mechanistic product-based solutions. While most of the

The implications of green roofs implementation within the capital is having an effect beyond the capital, in particular in the Greater London Watershed – particularly the area known as the Thames

Gateway. This area includes much of East London and stretches along the Thames estuary into Kent and Essex. Already there are a number of huge developments, which are under pressure to include green roofs as mitigation and compensation for ecological reasons.

However, there will be a need to constantly refine policy to meet the changing climate and needs of London. Green roofs are certainly no longer a fad but are an ever-growing market and an integral response to

impending climate challenges. The hope is that policy makers and regulators ensure that the roofs that are installed are of the quality needed to meet the capital's diverse and evolving agendas.

Basel In Basel, one of the most important pioneer projects at the time was the construction of an additional tract of the City's University Hospital, which foresaw green roof areas on all new buildings of the hospital complex. Patients should have a more pleasing view over the roof landscape, which in turn would ideally have beneficial effects on their healing process.

Research on sites like the University Hospital focused the potential for urban wildlife based on bird studies. The interesting results showing how black redstarts benefit from green roofs created a link to Dusty Gedge, an urban wild life specialist and planner in London. Dusty Gedge visited Basel in 2000 in regard of Stephan Brenneisen's bird and other studies. The knowledge of the benefits for black redstarts (a national endangered species in UK, with only two populations left in London and UK) by green roofs was the base for Dusty Gedge campaigning green roofs all over UK later.

Energy Saving Fund and the First Green Roof Campaign in Basel To get to the comprehensive green roof research of the University of Basel (later Zurich University of Applied Sciences, Wädenswil) the city of Basel did important contributions. In the early 1990s the city of Basel implemented an ecological mile stone with a new law to support energy saving measures. According to this unique law for Switzerland some 4% of all customers' energy bills are put into the "energy saving fund." Out of this fund campaigns and measures to support energy savings can be financed since then. As one of the benefits of green roofs is improving insulation and reducing energy consumption of buildings, a campaign for green roofs could be launched in 1996 based on the energy saving fund.

Following the "Year of Nature" in 1995 the project called "the improved roof" combined energy saving issues with nature conservation and got a total sum of one million Swiss Francs to subsidize house owners building their own green roofs. Within 18 months, over 120 green roofs were built in Basel, an area equal to 8 football fields. The project was carried out in cooperation with the trade association, which

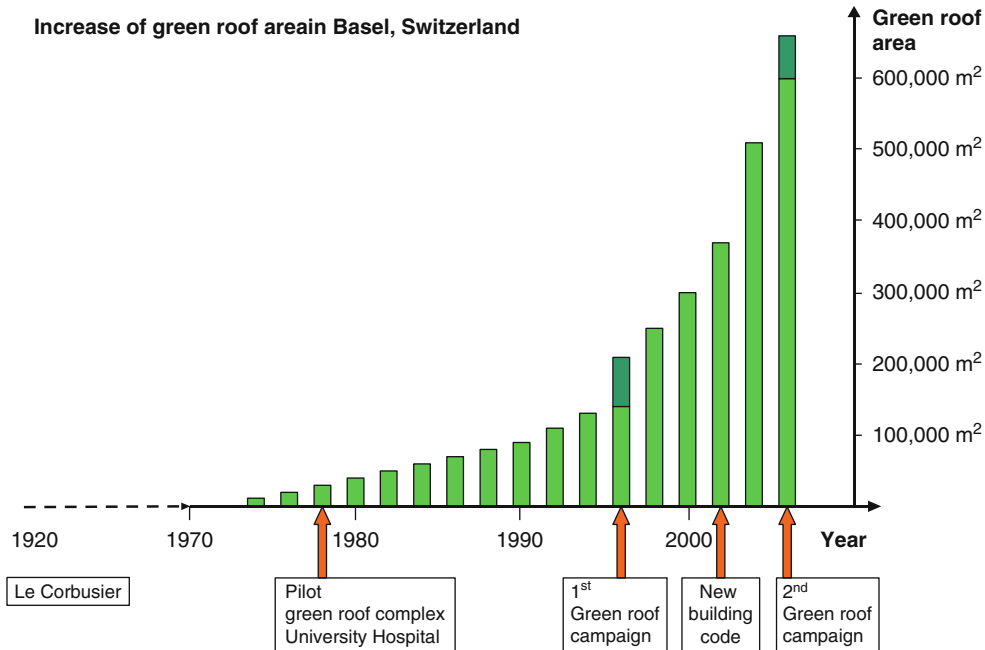
enhanced local know-how in the field. During the campaign, architects, planners, and the contractors installing the green roofs could get good practice and experience in this new technology – an important step to bring the measure into practice later.

Building Code and Wildlife Potential After the first green roofs campaign, Basel Canton passed an amendment to its building code (paragraph 72) in 2002, requiring all new buildings with flat roofs to have green roofs. This amendment was made based on the various benefits and impacts green roofs have for the urban environment and not least in recognition of the potential for green roofs to support biodiversity and species conservation in Basel. It is supported by additional specific guidelines on implementing green roofs in Basel to maximize their nature conservation potential.

The recognition that green roofs provide valuable habitats and support nature conservation objectives was later one of the drivers for Basel's second campaign, which started in 2005. This second campaign funds both green roofs and roof insulation. It was funded in the same way as the first campaign with a total sum of SFr. 1.5Mio.

Quantifying Green Roofs: Urban Heat Island, Climate Change, and Global Warming After the second campaign, again a survey recorded by aerial view photos the actual number and area of green roofs in the city of Basel. As a result 1,711 extensive green roof projects and 218 intensive green roofs (roof gardens) could be registered. So, approximately 23% of Basel's flat roof area is installed with green roofs in 2006 (Figs. 2 and 3). This amount will clearly have an increasing beneficial impact on the city climate. Regarding urban heat island problems accentuated by global warming the green roof policy of Basel could be an example: how to face one of the possible measures with a campaigning strategy getting green roofs from pilot to mainstream successful.

Building Code and Price as Important Drivers for Success The general public in Basel still finds green roofs "special and exciting," but for developers, installing green roofs is now considered routine. The ongoing reduction of the installation prices of green roofs from SFr. 100.-/m² (US\$90.-) in the 1990s down



Green Roof Planning in Urban Areas. Figure 2

Increase of green roof area in Basel, Switzerland, from 1970 until 2007

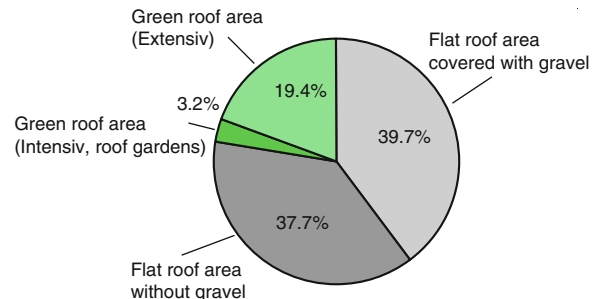
to SFr. 20.- (US\$17.-) was an important factor in the inclusion of green roofs into the building code (Fig. 4).

Green Roofs in the Rest of the World: A Summary

Green roofs as an urban planning tool is not becoming more prevalent throughout the world. Although the approach evolved in Germany, Austria, and Switzerland, it is now spreading throughout Europe and North America, and there is increasing interest in South America, Asia, and Australasia.

The oldest green roof policies were derived in Germany and Austria in the early 1980s. Stuttgart in Germany is probably the most renowned of the Germany green roof cities, although green roofs are an integral part of planning in the cities of Munster, Frieberg, Berlin, Dusseldorf, and Munich, to name a few.

In 1985, the municipality of Linz in Austria undertook an assessment of the loss of green areas and the quality of life in the built environment through a unique mapping process, known as the Green Space Plan. Although there were a number of green roofs within the city's boundary, including one of the largest



Green Roof Planning in Urban Areas. Figure 3

Percentage of green roof area related to the total flat roof area in Basel, Switzerland

in Austria – the Schachemeyer Factory green roof, this research demonstrated the need for a more direct approach to encourage more uptake of green roofs. The research became the basis for legally binding building plans, which included an obligation for building green roofs.

Green roofs were seen as effective solutions to “greening” in areas of Linz where land use was not compatible with open space development. This was



Green Roof Planning in Urban Areas. Figure 4

particularly important in the city's commercial and industrial zones. As a result, green roof policies were introduced in Linz in 1985 as legally binding and compulsory building plans. The direct impact of this strict measure led to an increase in green roofs, especially in the industrial areas close to the River Danube.

In 1989, the city of Linz instigated a generous financial incentive for building owners, by sponsoring green roofs by up to 30% of their, which was reduced in 2005 to 5%. Over the past 20 years, all these measures have resulted in over 404 green roofs being installed in what is by all accounts a relatively small city. The total area prior to 2007 was 400,000 m² (equivalent to 40 football-pitches). About 90% of this area is extensive green roofs and 10% intensive green roofs. However, the recent covering of the main motorway through the city with an intensive green roof (which took all the city's annual green roof budget increased the total area by at least 40%.

In North America, many cities are now actively promoting the use of green roofs. The first city to undertake this process was the city of Portland

followed by Chicago. Other cities include Seattle, Vancouver, Minneapolis, Toronto, Washington DC, and New York.

In Asia, the standard bearer for green roofs has been Singapore, which has been actively promoting green roofs for nearly 10 years and there is increasing interest from China and India in how green roofs can address urban health and environmental issues.

Bibliography

1. Grant G (2006) Green roofs and green facades. BRE, Bracknell
2. Hyder Consulting (2006) Green roofs for eastside outline drainage policy
3. Knaupp W, Staiss F (2000) Photovoltaik – ein leitfaden für anwender. Fachinformationszentrum karlsruhe
4. Kohler M (2002) Photovoltaic panels and green roofs – positive interaction between elements of sustainable architecture. In: RIO 02 world climate and energy event, Rio de Janeiro
5. Kohler M, Malorny W (2006) Wärme – dammeigenschaften von dachsubstraten mit vegetationsschnitt extensiver grunddächer. *Dach Grün* 15(3):8–13
6. Liu K (2002) Energy Efficiency of and Environmental Benefits of a roof top garden. National Research Council Canada, Ottawa. www.professionalroofing.net

Green Roofs, Ecological Functions

MANFRED KÖHLER¹, ANDREW MICHAEL CLEMENTS²

¹University of Applied Sciences, Neubrandenburg, Germany

²Green Roof Greece, Athens, Greece

Article Outline

Glossary

Definition

Introduction

Future Directions

Acknowledgements

Bibliography

Glossary

FLL The Landscape Research and Development Society (FLL) nonprofit organization was founded in 1975. Its mission is to research, produce, and disseminate all the various landscape development principles, guidelines, and specifications for the assurance of environmental quality [1].

FBB The Green Infrastructure Association (FBB) is a specialized group that was founded by some members of FLL to focus more specifically on green building. The FBB is the German counterpart to the American industry association Green roofs for Healthy Cities (GRHC) and one of the founding members of the World Green roof infrastructure Network (WGRIN). The German Word “Bauwerksbegrünung” has no translation in English – Green infrastructure in the sense of FBB is focused on all forms of urban green.

Extensive green roofs (EGR) also called natural green roofs, or eco-roofs, are vegetated roof constructions that require low maintenance. Drought-adapted plant species are used to create a self-sustaining vegetated surface suitable for nearly all types of buildings. Growing media is about 10-cm, or 3-in. thick [2]. The term “Natural Green Roof” is an own further term, which should set the main focus on enhancing the biodiversity on vegetated roofs. This can include in some regions irrigation with rain or gray water. Natural green roofs means engineered

green roof systems under the guidance of nature conservations solutions. This term includes extensive and intensive green roofs.

Intensive green roofs (IGR) also known as roof gardens are garden structures on top of buildings and other artificial urban surfaces. In most cases, the growing media is more than 20 cm deep. For trees, it can be more than 1 m. IGRs, with structures including lawns, planter boxes, shrubs, and small trees, require the same maintenance as traditional gardens.

Storm water runoff Rain water running off impervious surfaces.

Green infrastructure overall phrase in the North America for all types of green roofs technology and other types of greenery on buildings, like vertical greening, living walls, and indoor greening systems. Green infrastructure in a wider sense includes photovoltaic technology and rainwater management.

Growing media engineered substrate for green roof purposes. Green roof substrates typically have low nutrient content and high drainage rates. Typical materials are expanded slate, shale, pumice, or recycled products.

Green infrastructure A term to describe the range of materials and technologies used to enhance urban environments. In addition to green roofs, this term also encompasses other related systems such as vegetated facades with climbers or living wall systems, indoor greening systems, rain gardens, photovoltaic systems, and other technologies. Roof greening can be combined with living walls, indoor plants, and ecological landscaping to enhance the built environment. The USEPA refers to structures specifically intended to manage wet weather as green infrastructure.

Leadership in Energy and Environmental Design (LEED) This is a US-based rating System by the US Green Building Council (USGBC). Benchmarks focus on energy savings, water efficiency, CO₂ emissions reduction, improved indoor environmental quality, and stewardship. The categories of achievement are: silver, gold, or platinum. In Australia, a similar rating system uses “stars.” After an extensive debate about the merits of such certification systems, Germany set up one in 2009. Certification can be an effective type of marketing; however, one

critique of existing systems is that there is not enough weight placed on vegetation.

Low impact development (LID) A storm water management and site-design technique to mimic the situation before construction of the settlements. Water usage, evaporation cooling, and water storage and drainage are such benefits of green roof infrastructures.

Definition

Green roofs are engineered constructions that include environments suitable for well-adapted plant species. In most cases, these types of roofing have a longer lifespan than conventional roofing surfaces. The following elements are built on top of the roof structure

- The underlying protective layer is made of an impervious material such as bitumen, rubber, polystyrene, or other similarly adequate technical materials, in short: roof protection membranes.
- Additional, root barrier layers are available to prevent the root penetration of lower layers. These are known as separation fabrics or geotextiles.
- This is commonly followed by a separate water-retaining layer, which could be a natural porous stone material or an artificial retention mat; in short, this is a drainage layer.
- On top of this layer, a filter fabric separates the retention layer from the next layer: the growing media.
- The growing media is, in most cases, a specially mixed lightweight soil material with well-selected components for storing rain water. Growing media are mixed for different purposes (for example, extensive green roof growing media differs from roof garden media in nutrient and humus content). Intensive roof garden growing media differs in that in the upper levels there is a higher content of humus and on the lower levels lower humus content.
- The vegetation layer can range from a shallow layer with mosses and only a few taller plants all the way up to full-blown roof gardens. As such, green roof maintenance requirements can be as little as an annual inspection or as much as is usual for ordinary gardens. The success of the vegetation layer depends on the careful selection of the other green

roof layers. For example, if the goal is to plant trees on rooftops, a special combination of all these components is necessary.

The maximum weight of the construction must be calculated carefully. On average, it varies between 40 kg/m² (8.33 lb/ft²) and can rise upward to about 350 kg/m² (71.7 lb/ft²).

The longevity of green roofs depends on whether they can be easily accessed with the basic equipment needed for the success of the project as well as repairs. Maintenance and repairs must be planned carefully. Certain areas of the roof like, edges and places around roof fixtures like skylights or climate control systems can be prone to structural damage. Architects follow design visions, green roofs on “unusual” and for plant development difficult situations. Some times it is too high or too steep for a suitable plant growth or for the needed maintenance work. The technical and biological limits challenge green roof professionals. It is the duty of green roof specialists to figure out the limits, which are set. Uncontrollable aspects of local climate, such as wind, temperature, and the intensity of storm events, can set limits to what is feasible from an architectural design perspective and ideas about developing living surfaces on roofs and facades.

Green roof technology has become internationally famous in the past few years. The books recommended at the end of this chapter offer a wide range of basic information from various regions in the world.

Introduction

The global human family is currently undergoing a subtle, revolutionary paradigm shift away from an unsustainable, industrialized, mass, homogenized, synchronized society to a sustainable, custom, micro, niche, bio-diverse world. This is seen in every aspect of human life – from the work to entertainment. Gone are the days when work was done 9–5, mass entertainment was enjoyed all together and gone is the factory fortnight in August where all headed for warmer climes for our 2 weeks of holiday. Today it's work and play around the clock telecommuting, working from home, downloading YouTube videos and interacting on Facebook and Skype. This global shift to an information-rich, knowledge society is mirrored in how the urban greening industry worldwide interprets itself.

Once upon a time, not long ago, a green roof predominantly meant an amenity space with garden furniture and well-manicured lawns and water features or maybe at its ecological best mass-produced Sedum mats laid over plastic drainage boards and geotextiles. With deeper movement into a sustainable concept of our lives, this has been replaced by ecological sensitivity, biodiversity concerns, water management and conservation, and other such similar ecological, sustainable ideas. This entry will discuss this shift in detail and what exactly it means.

Green Roofs as Ecosystems

In recent years, there has been a growth in interest in natural green roofs, or eco-roofs, as distinct from roof gardens due to the ecological benefits that may be derived. Roof gardens were traditionally installed in many countries to provide amenity space and for building beautification, in places such as hotels and resorts. The ecological benefits, such as thermal insulation, that were derived were almost an afterthought. Nowadays, in many instances, natural green roofs are installed to become natural ecosystems in urban areas. The benefits achieved by such installations can be measured not only in the building itself but the surrounding urban area as a whole.

Green roofs have been shown to mitigate the urban heat island effect. Two key studies in the hot summer climate of Athens [1–3] concluded that the thermodynamics of both the building and the surrounding area are effected by the natural green roof installation studied on the Greek Treasury in Constitution Square in Athens. In addition, leading natural green roof researchers have compiled a review article, which explains this technology [4]. The concept of natural ecosystem green roofs may have grown out of the idea of just letting pioneer plant species grow on built surfaces in urban areas. Certainly, where this idea originated is a moot point now as it has become an emerging architectural style, which is being used by leading name architects with global reputations.

Natural green roofs are also called bio-mimicry or bio-phillic architecture [5, 6]. Ecologically, natural green roofs can be compared to natural structures such as rocky outcrops or cliffs where harsh weather conditions and shallow layers of substrate are found.

It is understandable why cities are sometimes described as being “urban canyons.” Many highly specialized ecosystems thrive in such structures [7]. In fact, these ecosystems, in the natural world, have often developed because of these extreme environments.

A number of environmental scientists and thermodynamics researchers are working on quantifying the energy savings, emissions reductions, and water resource management benefits that come from the use of natural green roofs. In addition, research is being conducted to quantify the many other ecological other benefits such as the reestablishment of plant and animal biodiversity in urban areas. Green roofs can be considered as the open ecosystems described by early ecologists [8]; where inputs of water, nutrients, and substrate to such installations occur from precipitation, dust in the urban canyon, and gaseous pollution. Losses to a natural green roof can come from wind erosion and storm water runoff. In terms of ecological productivity, natural green roofs have been compared to desert ecosystems, where there is a similar input/loss situation.

The total phyto-mass of a natural green roof is measured by considering the total shoot/root biomass. This figure varies in Europe between 100 and 500 g/m² total dry organic matter [9]. The 500 g/m² is an almost fully covered natural green roof, which has been planted with Sedum or similar succulents, Chive (*Allium spec.*), wildflowers, pioneers, and grasses with about a 95% coverage and about a height of 25 cm. The figure for a south-facing roof with coverage up to 60% is about 100 g biomass/m². The gaseous exchange (O₂, CO₂) productivity of such ecosystems is relatively low when compared with richer ecosystems such as forests. A biomass survey found a natural green roof with a 168 g/m² above ground biomass and about 107 g/m² root system biomass 2 years after installation [10]. These recent dry matter values are similar to the above-mentioned German measurement from the 1980s. If the main aim of the installation of a natural green roof is CO₂ sequestration, annual cutting of the plants is recommended for mulching, followed by allowing the plants to grow back each year. Natural green roofs are also being seriously considered to not only supplement urban agriculture to supply cities, particularly in the third world, but also in the West, as urban agriculture can lead to significant emissions

reductions. This is due to the fact that food can be sourced and consumed locally, reducing the need for transportation from rural areas into cities. There are notable examples of this emerging concept abound, across the world, from New York City to Bangladesh. Urban agriculture is substantially more productive using high-tech hydroponics and other such intensive technologies than it is with natural green roofs. So, green roofs can be used to supplement urban agriculture. In addition, natural green roofs can provide specialist agricultural products such as aromatic herbs for culinary and pharmaceutical use.

Natural green roofs are usually designed to keep vegetation conditions stable over the long term. However, there is a growing belief, particularly among natural green roof pioneers in Greece, that natural biological succession could result in the development of highly successful, bio-diverse, specialist ecosystems, which have adapted to the peculiarities of urban centers. On the whole, human intervention and “maintenance” should be kept to a bare minimum, if not removed altogether. It should be noted that it is advisable that a specialist inspect a natural green roof annually to check the waterproofing and to make sure that the vegetation has not become dangerous in terms of height. Apart from that, nature should be allowed a free reign. It must be noted that natural green roofs are not usually designed as recreational space. Furthermore, safety concerns relating to the height of the vegetation must be addressed. Barriers to prevent falling of material must be installed around natural green roofs.

Natural green roofs, then, can be considered extreme man-started natural habitats on artificial urban surfaces. After the installation annual inspection means that these structures require low to nonexistent human intervention. Environmentalists are becoming increasingly interested in the almost self-sustaining nature of natural green roofs. Overall, researchers are beginning to realize that they may well be able to reduce the anthropogenic impact of urban environments on the biosphere. The level of this impact depends on a plethora of factors, yet, these can be engineered by the design of the structure.

In this entry, three main aspects of natural green roofs will be highlighted. First of all, the benefits of natural green roofs will be discussed. To assess this, a description of natural green roofs as ecosystems will now be made.

Green roofing is currently (2010) practiced in about 40 countries worldwide please see www.worldgreenroof.org. There are organized green roof associations in about 20 countries at the time of writing [11]. Also, about 500 peer-reviewed articles have been published concerning green roofs at the time of writing this article. This would suggest that while green roof technology appears to be quite simple and low tech, the concept, particularly natural green roofs, is in fact complex [12]. Much more research is required and will be conducted going forward.

What is required is a comprehensive picture, based on research, of what exactly are the benefits of natural green roofs. It is certain that all forms of green roofs and green walls will play a fundamental role in terms of thermal insulation for buildings and for mitigation of the urban heat island effect in the future.

It seems certain that the performances of natural green roofs will aid a revised perception of what is aesthetically pleasing toward natural green roofs and away from traditional roof gardens. The new perception will be that bio-phillic is better than manicured human-created gardens [13]. The description of the ecological functions of natural green roofs by pioneers, such as James Todd [14] and his more modern counterparts such as Peck [15] and green roof enthusiasts, campaigners, and activists like Dusty Gedge in the UK (www.livingroos.org), helps to create new paradigm in what a green roof should look like. When the thermal insulation performances of natural green roofs are combined with water conservation and enhanced biodiversity on a massive scale around the world, a technology emerges, which could ameliorate, if not solve, some of the most pressing concerns of our time. This technology may soon become the status quo in architecture.

One of the main challenges of the urban greening industry and its stakeholders now and in the future is to obtain a full and detailed understanding of a natural green roof ecosystem, its energy cycle, and how these systems interact with other green technologies. It must be stated that there are already peer-reviewed articles, which begin to provide insight into this complexity and could make substantial contributions to architectural textbooks in the coming years.

It should be pointed out that roof gardens or intensive green roofs, which are primarily constructed as

amenity and recreational space in towns, could also be used for ecological concerns. For example, if roof garden decks are built with rainwater catchments and rainwater storage systems, there are ecological benefits that can be achieved. The deeper substrates that are used in roof gardens, which often range between 50 cm and 2 m offer enough substrate depth for many species of shrubs and small trees. Of course, safety issues must always be addressed before planting taller plants because at roof level, falling plant is a serious concern. When done properly, urban forestry may be possible on roof top spaces. Susan Weiler's Church project [16] is one example of an attempt to achieve urban forestry. While this project is about traditional horticultural ideas of aesthetics, it does also incorporate the use of trees using forest mimicry. It must be noted, once more, that safety must always be put first when dealing with the urban environment. Which tree species have been considered as suitable for urban forestry? First of all, many common or ornamental plant species have been considered. All phyto-mass-related functions can be delivered. Research about the potential of such forestry structures on rooftops has been conducted and tested in Hong Kong [17].

One conclusion of this entry is that if the load capacity of a building allows greater structural weight, a deeper growing medium opens up a wider range of potential plants species for selection as long as the aforementioned issue of falling has been addressed thoroughly. Consequently, urban forestry becomes possible along with natural green roofs. The challenge for the urban planner is to find appropriate mixes of planting schemes and appropriate safe technologies for particular climatic conditions and building codes. In recent years, forestry structures that have been installed on buildings range from sparse plantings to full earth shelters with copse planting schemes. Greened buildings, then, can have significant thermal performances and also act as natural corridors for the reintroduction of nature into urban areas [18].

Introduction to the Three Main Benefits

Green roofs, and especially roof gardens, offer aesthetic benefits and urban amenity and recreational space for city inhabitants. The measurable ecological and economic benefits that are derived by urban greening make

these technologies increasingly attractive to town planners, architects, and civil engineers alike [19]. Roof greening serves both private and public interests. There are also local and global benefits.

Proponents of green roof technologies, such as The World Green Roof Infrastructure Network, (<http://worldgreenroof.org>) believe that all local and national governments should institute incentives to encourage the implementation of massive scale urban greening. The experience of German roof greening, in the last 30 years, suggests that government incentives to individuals, businesses, and the state lead to significant increases in the implementation of these technologies. This model can be copied around the world. To date, the cities that have taken these technologies most seriously resulting in large-scale implementation are mainly found in Germany. Cities such as Berlin and Stuttgart have high adoption rates, which are measured in percentage points of available city area. It must be noted that cities such as the Austrian city of Linz follow close behind. A full overview of the state of roof greening in Europe was given in a workshop at the World Green roof congress in Nürtingen 2009 [20, 21].

Some European Cities, such as Copenhagen, Malmö, and London, are jumping on the roof greening bandwagon with ambitious plans for roof greening in the coming years. An important next step in the implementation of large-scale greening globally is the mapping and modeling of successful initiatives that are being taken by local and national governments, agencies, and stakeholders.

The three main benefits that are derived by roof greening are the following:

1. The influence on the urban water management cycle
2. The energetic performance of green roofs related to the buildings, the city, and the planet as a whole
3. The development of urban biodiversity, leading to a richness of flora and fauna in urban areas

The Urban Water Management Budget The urban water management budget is stressed by the impervious surfaces of a built environment. In the natural world, storm water runoff forms a small percentage of the total water volume that falls as precipitation. Most of the water that falls as precipitation, in the natural

world, is absorbed by the Earth and becomes ground water. The reverse is true in the urban built environment: Most of the water that falls as precipitation becomes runoff. This causes a number of problems. Athens may be representative of what happens during a storm. The city squares are turned into lakes, and the streets become rivers within minutes of the onset of a storm event. Recent research into urban water management budgets conducted by Göbel et al. [22] suggests that urban landscapes could be transformed using a combination of rainwater catchment and retention areas, use and absorption, which would reduce urban water management costs to a level not much different from water management costs found in the natural world. In this context, green roofs are important as a type of decentralized rain water catchment and retention system. Green roofs and vegetated facades can retain storm water and release excesses over a period of hours, if not days. Green roofs act as a brake on runoff as they absorb and store large percentage of the water that falls as precipitation in a similar method to that which is found in the natural world.

Rainwater management with green roofs can incorporate storage values not only in the growing media but in all other layers in green roof systems. The drainage layer can be used for storage. Surface and subterranean ponding can be designed to allow further storage. In addition to the defined water capacity of the growing media, drainage/storage layer, ponding, in countries where monsoon and excessively heavy rainfall events are experienced, growing media can be designed and plants selected that can store even larger quantities of water. Again safety issues are an important consideration here because water is heavy, so permissible static loads must be taken into account, particularly in seismic regions of the world. Consequently, it becomes clear how runoff is reduced in quantity and delayed over time by green roofs.

When green roofs are combined with other rainwater catchment and retention technologies, such as rainwater harvesting, re-use, and infiltration, it is possible to create a zero runoff position in a city. The decentralized storm water retention property of green roofs also reduces pressures on wastewater treatment systems that are already overloaded. The peak runoff after a heavy storm event is minimized.

This ability of green roofs to catch and store rainwater also has an impact on the second major benefit of green roofs in their thermodynamic properties. Surface runoff rates and the lack of green spaces can explain why there is almost no evaporation and why urban climates can be hot and dry [23]. With the increasing number of megacities, this climate may contribute to health problems for many citizens [24]. As will be seen in the next section, green roofs are active thermodynamically. This sets them well apart in terms of their thermal performance from conventional forms of thermal insulation using materials like stone wool and polystyrene. This will be discussed further shortly. For now, it can simply be stated that the captured water in a green roof, during a storm event, is evapotranspired by the plants. This actively cools hot cities. This cooling effect reduces air conditioning demands. Reduced air conditioning use further cools the city, leading to a virtuous cycle. This will be discussed later.

In developed countries, in the next few years, sewer systems that are approaching the 100 years of age mark will need to be replaced. The decentralized water management nature of green roofs, which reduces storm water runoff quantities in cities can provide a viable solution to this issue. This has been demonstrated in Berlin [25]. Since the 1980s, numerous projects to reduce storm water runoff, manage water demands, and handle wastewater disposal in cities have been implemented in Berlin; 17 projects were documented, and a list of these projects is downloadable from the official webpage of the Berlin senate [26]. Green roofs and green walls are fundamental components of green building technologies in all of these projects [27].

Increasingly, over the years, peak runoff during heavy storm events as well as the “Urban heat island effect” has become ongoing and challenging problems in inner City water management. Open evaporative surfaces could well become mandatory solutions [28]. In Berlin-Brandenburg, about 80% of the total precipitation that falls is converted to evaporation using various technologies and methods, which include green roofs. This figure could become an achievable target for all cities on Earth. This also links to the third benefit of green roofs, which is the development of natural ecosystems in urban areas.

In Berlin, for example, for each type of inner city habitat, an optimized vegetation value can be

calculated. In Berlin, this value is referred to as the “biotope area factor.” These basic strategies must be adapted to suit regional climates with appropriate plants and growing media alternatives and local building codes. A further link between the benefits of green roofs is their effect on the so-called urban heat island effect, which links the water retention benefit with the thermodynamic performance. The urban heat island effect is partly caused by the almost nonexistent property of evapotranspiration in a city. By retaining more water in urban structures in green roofs, the city is cooled in the summer, mitigating this phenomenon. In addition, in an emerging paradigm among climate specialists, water cycles are now considered to have a fundamental causal relationship with man-caused global warming [29].

To develop large-scale vegetation structures on buildings, cooperation between architectural disciplines and green planners, researchers and designers must begin; the conventional so-called black roofers, who use bituminous and plastic membranes, and so-called green roofers, who use water storage/drainage systems, substrates, and plants, must engage in productive dialogue.

Currently, about 5–10% of new buildings in Germany have green roofs. The number of green walls remains, at present, less than 1%. This potentially represents a huge opportunity for environmentally friendly architecture. A paradigm change in architecture is needed to construct not only in terms of energy-efficient designs but also buildings that have zero negative impacts on the environment. This is the challenge for the coming years, and green infrastructure will play a dominant role in assisting cities to achieve this objective [30].

Energy At the dawn of the twenty-first century, humanity faces what could possibly become its defining moment. The choice is becoming ever more obvious, and humanity is going to be forced to make a decision soon, en masse, about whether humanity makes it to the twenty-second century or becomes just another flash in the pan for this planet. The human race potentially faces the horrific possibility of a massive die-off at the very least or extinction at its worst. The reason is quite simple: Mankind has been able to develop modern civilization on the Earth, and achieve

the undreamed of standards of living and technologies that it has due to the tremendous subsidy that abundant oil has provided. One liter of oil is the energy equivalent to 100 man-hours of labor. The West has collectively burned or used a large percentage of the oil reserves on Earth during the last 100 years. How much oil has been used and how much is left is open to debate and probably unknowable. More may be found somewhere, sometime. What cannot be debated is the fact that oil is a finite resource, which took millions of years to form, and if only the West continued using it at the rate that it does, it will dry up sooner or later. This is likely to be sooner rather than later because now the rest of the world is developing at rates that make Western industrial development in the late eighteenth and early nineteenth century looks like child’s play [31].

In addition, climate change, global warming, the coming ice age, how much influence human activities have on the Earth may be debated endlessly. What cannot be questioned is that continued production of CO₂, CO, and a multitude of other poisonous gases through oil combustion, ongoing quarrying, mining, deforestation, increased urban development, natural ecosystem degradation cannot end well for humanity or for most forms of life. It is inevitable that, again, sooner or later, the human species will have cut down the last tree, used the last drop of water, polluted everything, everywhere, and this is obviously not beneficial for life.

Certain issues require detailed, analytical, scientific study and understanding to attain a deep awareness. This issue can be understood by common sense alone, most effectively. If you keep taking something from a finite source, sooner or later, there will be none left. Moreover, if by using the finite resource taken, degradation of your life-support systems occurs; the end to this scenario does not require much thought to calculate.

So, humanity faces three of what are possibly the greatest challenges any species can face. The first is increasing rates of depletion of ever dwindling energy and natural resource supplies, the second a buildup of toxic waste in its environment, the third exploding population growth. Thankfully, a viable solution to these three scary challenges is available. While it is true that these points may seem exaggerated and the solution simplified. Increasing numbers of highly

acclaimed scientists from around the world are suggesting that the solution comes from the cause of the problem. Nature is the issue, and nature is the solution.

Some activists and campaigners state that there is a need to save the Earth. The Earth has been here, according to most estimates, for 4.5 billion years. In that time, it has survived far greater threats than humanity. Possibly, what humanity needs to do is to save itself. The only savior up to the task is nature itself. This is where urban greening and particularly green roofs really prove their value. Green roofing may well be the ultimate “return to nature” that was much touted by groups in the 1960s and subsequent eco-friendly parties and interests but this is a “return to nature” without having to move anywhere. Cities, can be taken, as there are and transformed into natural ecosystems using green roofs. The impact of this on the challenges would be, to understate it, revolutionary and evolutionary.

Possibly, the most serious and important impact of this would be its effect on human energy requirements. Numerous studies conducted around the world have shown that green roofs significantly reduce heating and particularly cooling requirements in both buildings and whole cities. In the aforementioned thermodynamic study of the oikostegi (natural green roof in Greek) installed at The Greek Treasury in Constitution Square, opposite the Greek Parliament in Athens, Drs. Rogdakis Ph.D., and Koronaki Ph.D., concluded that substantial thermodynamic effects were observed.

Air conditioning requirements were reduced by 50% in the floor directly beneath the green roof installation. This is important when one considers that only 52% of the roof surface is covered by the oikostegi natural green roof. A further important conclusion of the study was the so-called thermal lag effect of the green roof. Peak environmental temperatures were observed at 1 p.m., the building temperature peaked 30 min later where it was not greened, and the area that was greened peaked a full 90 min later at 3 p.m. This is rather important for two reasons. Firstly, due to the fact that only 52% is greened, there are large thermal bridges, which would not be observed had the entire surface been greened. Secondly, this particular building houses government offices and the peak demand for cooling of the building has been delayed by the natural

green roof until 3 p.m., at which time the civil servants have already gone home. This means that there are further reductions in air conditioning because it is no longer required. Before the installation, peak requirements were observed 90 min earlier. On the surface, this may not appear substantial but the Treasury has an air conditioned office space that totals nearly 12,500 m² <http://www.oikosteges.gr/index.php/greenroofs/research>.

Another important aspect of the thermodynamics of green roofs is the fact that because air conditioning use is reduced, further reductions in the ambient temperature of the surrounding area are reduced leading to even further reductions in cooling requirements. It was anecdotally stated by a Greek thermodynamics specialist that if all the air conditioning were switched off in Athens for 3 days, Athens would not need air conditioning. The reason for this is simple. The inventor of air conditioning, Carrier, never designed it as a method of cooling. It was designed to condition the air. The way air conditioning works is by taking hot air out of a building and dumping it into the surrounding area. This means that the surrounding area becomes hotter leading to an increased requirement for air conditioning leading to increased dumping, leading to a vicious circle. What green roofs do is to reduce air conditioning requirement, leading to falling ambient temperatures in the city, leading to further reductions in air conditioning requirements in a virtuous circle. Large-scale implementation of green roofs leads to a lowering of peak summer temperatures in cities, which leads to a lowered energy requirement to run air conditioning units and lowered peak ambient temperatures.

A similar story is observed in the winter with heating requirement reductions. This number of reduction depends on various factors, but the values vary between a few percentage up to 20% in roof top level apartments [43]. Reducing energy requirements with green roofs has another important impact, especially in countries like Greece and Germany where lignite is burned to produce electricity by the Public Power Corporations in both countries. Lignite burning produces even greater carbon emissions than oil burning and is much less efficient. Carbon is one of the main greenhouse gases, which are thought to be partially responsible for global warming. Energy efficiencies achieved by green roofs, then directly reduce carbon

emissions, reducing greenhouse gases and slowing down global warming. It must be noted here that green roofs also absorb greenhouse gases, further reducing existing emissions. The full impact of green roofs is still not completely known, and there is room for much research. Having said that there is little doubt that they have a substantially important impact on reducing energy requirements, which has an impact on reducing emissions and heat caused by air conditioning, which has a further impact on reducing energy requirements in a virtuous circle.

In Europe, the goal for fossil fuel consumption reductions stands at 40% by the year 2020. Furthermore, European countries have pledged to reduce carbon emissions by between 10% and 20% by 2020. Greece has ambitious plans to reduce carbon emissions by 30% by 2020. All of these goals become achievable using massive scale urban greening using green roofs. Green roofs can return their installment investment cost many times over in energy requirement reductions and emission reductions very quickly.

It should be stated here that green roofs operate rather differently from other forms of thermal insulation. A green roof is an active cooling and heating device as well as a superior form of thermal insulation. Conventional insulation isolates the environment from the building, keeping heat out of the building in the summer or by trapping heat inside the building in the winter. The way it works is by creating a thermal barrier between the building and the outside environment, which prevents thermal flow. A number of materials are used, including stone wool and polystyrene.

A green roof insulates a building in this way, but it also plays a more active role in the thermodynamics of the building and, in fact, the surrounding environment. First of all, it acts as a thermal pump in the summer. Heat that is in the building is drawn out by the green roof as the green roof temperature is lower than the environment and the building. Also, a green roof lowers ambient temperatures in the surrounding area. So, if green roofs are installed on a large scale, ambient temperatures would fall in the whole city in the summer. The same is not true for conventional thermal insulation. Moreover, as stated earlier, as ambient temperatures fall, air conditioning use falls leading to further falling ambient temperatures. Again, this only applies to green roofs.

Evaporative cooling consumption by green roofs is the name given to these processes. Further improvements to the performance of green roofs can be made if rainwater catchments and green roof irrigation are utilized. This is known as adiabatic cooling, which is a low energetic cooling process in which rainwater can be used.

If roof surfaces are extensively greened, storm water runoff can be detained and retained [28] to irrigate the green roof during the summer to improve the thermal performance still further using evapotranspiration.

So, the evaporation of water in a city is an important component to reduce the heat impact of solar radiation in a city. Lower evaporation rates mean higher surface temperatures and in urban areas a main contributor to the “urban heat island effect.” On a global scale, the reduction in evaporation in the built environment is mainly responsible for climate change [29]. Mitigation of global warming by evaporative cooling is one key process, which can be used in the reduction of global warming. A special entry on this was delivered at the climate conference to Copenhagen by Kravčík et al. The conclusion of the entry is that global warming is related to the loss of vegetation worldwide. Green roofs offer new space for vegetation structures and provide the needed evaporation cooling systems.

Biodiversity Urban biodiversity is the third benefit that green roofs provide. The Convention on Biodiversity (compare: <http://www.cbd.int/cop9>) COP9 Declaration of the Congress in Erfurt in May 2008 demanded that green roof plant selection foster and encourage urban biodiversity. Environmental laws, such as environmental building codes, help to mandate these demands at the national level. Planting trees wherever possible in a city is the first step in solving the issue. This is possible on parking space areas in the commuter zones of cities; tree planting in cities also reduces the air conditioning requirements in cars that have been in the shade, when driving recommences. Parking decks can utilize shade trees to protect cars from direct solar heating. This reduces air conditioning requirements in cars, which again feeds into the virtuous circle mentioned earlier. Parking decks are an important source for greenery. In Singapore, for example, there is a government program to set up green roofs in parking areas as an open space resource.

Most natural green roofs (NGRs) are promoted as self-sustaining ecosystems. A plant cover of about 60–75% indicates that a high-quality green roof has been achieved; however, in many cases, the focus is only on few specific types of plant species. NGRs can and should also be designed to encompass a high range of biodiversity. Depending on the type of the project, a focus on regional native plants can be integrated into the design of the installation. To achieve this end, the right selection of plant species, proper design of substrates, and a variety of different methods of ecosystem establishment are important.

The number of plant species on a green roof depends on a number of parameters, such as the manner in which plants are introduced to green roofs. The following is a brief description of the most common methods of plant establishment:

- Seeding (affordable for flat roofs).
- Sedum sprigs (cheap and easy, establishment takes about 1 year).
- Plugs – there are a nurseries, in countries where roof greening is popular, around the world, that specialize in green roof plants.
- Preproduced turf mats. This technology was developed a few decades ago. Most of the turf mats include drought-adapted grasses and also several *Sedum* species. Depending on the local situation, grasses are often the best competitors under shady conditions, and *Sedum* species are usually the winners in the areas with full sun. This method offers a complete vegetation cover from the first day of installation. This is important for pitched green roofs and high-rise buildings with a high wind uplift to avoid the erosion of soil and decline of plant species.

Seasonal changes in the vegetation structure can be observed. The timing of blooming on extensive green roofs is related to the frequency of rain events. In humid and wet summer seasons, more flowering is normal; also more pioneer plants invade spontaneously during wet seasons. The acceptance of unplanned plant species depends on the aims of the project.

Weed Control Some plants are known to be difficult plant species for NGRs. These must be removed as early as possible. An annual inspection and weeding can be helpful. Also, it must added that this is an issue of great

debate in the green roof world. Weeds can be defined as plants that grow and are not wanted. Some so-called weeds can be invasive and dominate if they are not controlled, which may or may not be desirable. Another way of looking at a weed is in terms of biological succession. Weeds are really pioneers. They are plants that take over barren desert like environments to begin to nature's slow process back to a climatic ecosystem. Nature is innately wise and so if the building owner, the architect, and the green roofer are aware, then weeds can be viewed as pioneers and a desirable addition to a natural green roof.

Water Management by Green Roofs, an Overview

Approach The most desirable situation in terms of water management, in a city, is to achieve zero runoff and 100% evapotranspiration of precipitation. Green roofs detain and store water by absorbing it in the ways mentioned earlier. This stored watered is then evapotranspired by the plants. The rate of evapotranspiration on a green roof is dependent on the amount of water captured by the NGR.

In order to calculate the various components of the NGR water budget, meteorological data are needed. It is easy to measure precipitation (normally with tipping buckets). Runoff can also be measured quite easily by installing adapted tipping buckets in the downspout. The water content stored in NGR substrates can be measured either with a type of roof lysimeter, or it must be calculated via a typical water budget equation, like HAUDE or others (Figs. 6 and 7).

Currently, about ten research institutions are working on this question to get realistic numbers on the retention rates by measurements on natural green roofs. Such experiments are being conducted globally and on different scales. Scale ranges from the module scale of a small research field, to measurements on small newly constructed test houses of a few square meter, like in PennState State researches in the US or measurements on realistic research roofs, to even larger study areas, like in Neubrandenburg or Berlin [31]. Since 2003, a green roof research facility has been collecting similar data in the British Columbia Institute of Technology (BCIT) in Vancouver [32].

This data about how green roofs perform is a major contribution to “applied ecological research.” Other

working groups are using this data to simulate the effects of NGRs at city blocks, neighborhood, and on even larger scales.

To test the performance of green roofs materials, like the drainage layer and growing media, the FLL [34] developed physical test procedures. These tests were carried out in replicate; however, they were only performed on the growing media without a vegetation layer. Research facilities on a real roof situation perform more realistically than test plots at research facilities. Green roof plant species need realistic harsh weather conditions of wind and drying out of growing media [33]. Thus, these values are comparable to but lower than real measurements done on real roofs.

The retention values are counted in different ways [33]. Compared annual retention rates of 12 different research institutions from Germany, Sweden, Belgium, and USA were made. The annual retention rates ranged between 42% and up to 80% on flat NGRs. For sewage calculation, the peak load capacity is the most important value. To count this, a technical report is included in the FLL guideline on how to make replicable measurements [34]. In Germany, all growing media must be certified with this value.

The effect of a pitched roof has no significant influence on the percolation of water, this was another early finding on German test plots situated in Hannover [35].

An Italian group [33] carried this idea further and investigated it in Mediterranean climate over a duration of 18 months' rain events with varying intensities. If the total rain event was less than 10 mm, during periods of good ground coverage (75% +) by the NGR, nearly 100% was retained. During the heaviest rain event of 132 mm in November, only 10% of the precipitation was retained; however, the peak flow reduction was nearly 80%. The time lag from the beginning of the rain of this major rain event and discharge was 148 min compared to almost immediate discharge on an exposed roof. This braking effect of an NGR is most significant for flood control in a city.

An interesting observation is the differing data requirements between professions, such as ecologists and civil engineers, who are responsible for drainage systems of cities.

The debate about NGRs rainwater retention benefits has two aspects:

- How can water be captured by the vegetation structures as well as in a storage tank? Is irrigation also a suitable solution for NGRs? After all, irrigation does help to reduce high temperatures, particularly in the late afternoon hours in summer. How can the evaporative cooling effect be measured?
- The integration of green roofs into a rainwater management strategy to achieve “zero runoff” from a property is desirable. This would also help to reduce mixed sewage overloads of inner city sewer systems. A combination of green roofs, artificial ponds, rainwater harvesting, and infiltration could indeed result in a “zero runoff strategy.” Local solutions are needed to achieve this goal for each property and for individual regions. Green roof technology is a “decentralized” technology. Political instruments, like the “runoff tax,” in Germany, could be helpful in establishing these strategies in city policies [36, 37].

Evaporation takes place for several days after a rain event on a NGR. The actual evapotranspiration amount depends on the amount of stored water in the substrate, the temperature, the air humidity, solar radiation, and wind speed. The amount of real evapotranspiration water on a NGR in central Europe varies in summer between 0 mm/day (dried out growing media) and 4.5 mm/day in saturated conditions in Neubrandenburg. In most cases, on extensive green roofs in Germany, there are drought conditions [38].

The measured evapotranspiration, by a agricultural field lysimeter station in Berlin, can be up to 8 mm/day. The potential evaporation on green roofs is high due to high solar radiation rates on a roof, high wind speeds, high temperatures, and low humidity in cities.

The full year measurements allow for selection periods with interesting parameters, the following case studies show the range of experiments that have been conducted:

- (a) Summer, dry or wet growing media – a light rain event
- (b) Summer, dry soil, heavy rain event
- (c) Summer, long wet period, effect of rain events
- (d) Winter, wet growing media, no frost some rain

- (e) Winter, wet growing media, frost, snow layer
- (f) Winter, wet growing media, frost, snow melt, more rain

When such periods are selected from the data set, what happens?

- (a) In the summer situation, nearly all rain quantities can be retained and be evaporated over the following days.
- (b) Surface runoff is lower if the vegetation cover is complete and the green roof is not inclined. On flat roofs, in normal situations, the rainwater will be retained in the coarse structure of the soil and by some ponding on the roof. In experiments, the quantity of the runoff was reduced and the time lag between the rain event and discharge was increased, as already mentioned, when a green roof was installed [39].
- (c) The diploma work of Bustorf [40] recorded during a 10-mm precipitation event in the summer in a wet period. A 50% storage capacity in the growing media, with the additional 50%, i.e., 5 mm, being detained for a number of hours and discharged slowly.

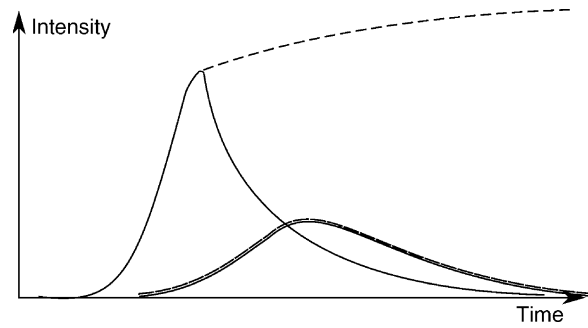
Various research groups from Neubrandenburg, Berlin but also in other climate regions have found remarkably similar results.

Figure 1 shows the typical structure of a rain event. Peak intensity is achieved rapidly followed by a much longer duration with lower intensity.

Green roofs arrest discharge for an extended duration, which, in most cases, is as long as 15–20 min after the onset of the rain event. On an exposed roof without a green roof, discharge is instant and complete. If some extra retention technology is incorporated into the green roof system, this time lag can be extended indefinitely. In terms of urban water management, time lag is critical. All impervious services, as is known, and as has been stated, have instant and complete runoff. That is why city streets become flooded during a storm event.

In addition, as already stated, discharge is also arrested and runoff is always much less on a green roof, making it easy for aging urban drainage channels to accommodate the minor amounts of runoff.

Table 1 shows evaporation capacity of NGRs without irrigation in the Northeastern German climate.



Green Roofs, Ecological Functions.

Figure 1

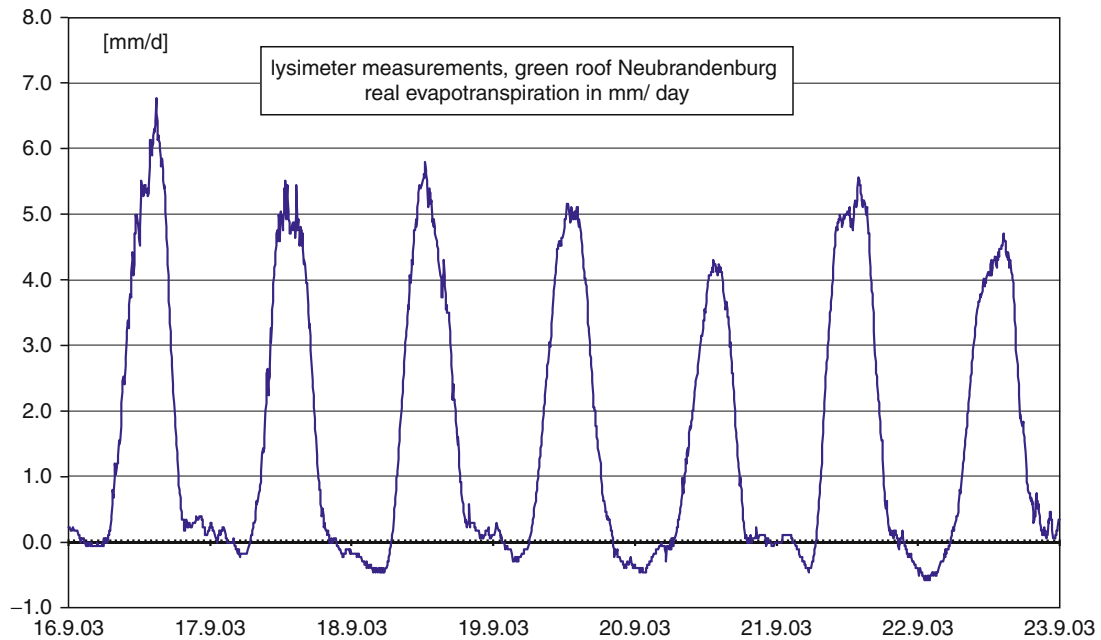
Schematic function of green roof retention values.

Typically, peak rain full quantity is achieved shortly after the onset of the rain event followed by decreased rainfall quantities over time as the event develops. In layman's terms, a short downpour is usually followed by subsequent ongoing drizzle in many if not most storm events. Consequently, green roofs are able to absorb the initial peak and then break the subsequent runoff occurring from the ongoing drizzle. In conclusion, the function of a green roof is to absorb initial storm event precipitation, store the water, and break the runoff of subsequent rainfall. The stored water is then evapotranspired over a matter of days and maybe even weeks after the rain event

Green Roofs, Ecological Functions. Table 1 Daily evaporation rates during different seasons (own lysimeter measurements, Neubrandenburg) [41 Koehler 2004, updated)]

Season	Evaporation values in (mm/day)	Growing media saturation
Winter	0.1–0.5	Well saturated, no frost
Spring/autumn	0.6–1.5	Well saturated
Summer, hot	1.5–4.5	Well saturated
Summer, hot	0.0–0.2	Dry substrate

The results of lysimeter measurements demonstrate that evaporation also occurs in winter. Summer evaporation rates depend on the amount of water stored in the growing media. Three millimeters per day of evaporation is the mean under saturated conditions (Fig. 2).



Green Roofs, Ecological Functions. Figure 2

Real evapotranspiration of a green roof in Neubrandenburg under saturated conditions, measured by lysimeters (compare Fig. 7)

The Table 2: shows the real evapotranspiration of a green roof lysimeter in mm/day. At night, the condensation process can also be observed. The evaporation of 1 mm of water means 1 L/m² converts 680 Wh of energy into latent heat. The resultant cooling rate of 3 mm/m²/of evapotranspiration in summer represents 2,040 Wh.

After more than 10 days without rain, green roofs, which have not been designed for extended drought show evidence of stressed conditions due to the lack of water in the system. Specially designed natural green roof systems, such as the Greek oikosteges natural green roof ecosystem, have been able to survive for over 6 months without any inputs of water. So, while the evaporative cooling capacity of a green roof may decrease the dry growing media, it still acts as a thermal insulation layer. This means that the green roof medium provides good insulation but is not always beneficial to the plants unless the whole system is purposely designed for such situations. It is the designer's responsibility to manage this conflict, by installing an irrigation system, accepting drought in the vegetation or designing systems for such situations.

Under normal cases, most green roof plants regenerate again, when it rains.

In most models and simulations of the potential daily evaporation rates of NGRs, rates are usually taken from calculations from data provided by standard meteorology stations, which focus on agricultural crop production. The difference with extensive green roofs is the vegetation's ability to survive quite long periods of water stress and regenerate like desert plant species. *Poa compressa* is an example of a plant that has this ability. If the goal is to create a constant cooling system with green roofs, then irrigation is advisable.

Studies measure the long-term retention by an extensive green roof. In Hannover, the effect of distance from the nearest drain to the point of study on retention and discharge rates was studied. The experiment systems studied were designed according to FLL green roof design guidelines. This study showed that discharge rate and discharge volume were dependent on the depth of the growing media, total green roof area, the roof inclination, and the distance to the nearest drain. The discharge values differ between 0.6 for very shallow substrates of about 4 cm and 0.4 for 15–20 cm

Green Roofs, Ecological Functions. Table 2 Energy demands of evaporation of rain water on different roof surfaces mean daily values [38, 41]

Roof surfaces	Net Radiation Wh/m ² /day	Latent heat flux Wh/m ² /day	Sensible heat flux Wh/m ² /day	Bowen ratio ^c
Mean global value	2,463	1,888	575	0.30
Bitumen Roof ^a	1,950	123	1827	14.85
Green roof ^a	2,057	1,185	872	0.74
Gravel roof ^b	2,132	687	1,445	2.10
Green roof ^b	1,800	1,258	542	0.43

^aUFA-Fabrik Berlin, 4th June until 31st August, 2000. Precipitation 201.4 mm. Mean global radiation 5,354 Wh/m²/day

^bNeubrandenburg, 26th April until 8th July 2004. Mean global radiation 4,848 Wh/m²/day

^cBowen ratio: ratio of sensible to latent heat in this context, when the magnitude of this ratio is less than one, a greater proportion of the available energy at the surface is conducted to the atmosphere as latent heat than as sensible heat, and the converse is true for values greater than one

depths under controlled conditions in green houses. The substrates were filled into boxes, which were then fully saturated with water before being subjected to periods of experimental rainfall. This enabled different growing media to be compared. The influence of plant species was not tested by this model.

These measured values can be used to calculate the retention values of typical green roof projects in comparison with standard roofing systems.

In Block 6, in Berlin-Kreuzberg, a system was developed to collect gray water and rainwater from 120 apartments to be treated in a constructed wetland and used to flush toilets. This reduced the demands on potable water by up to 50%. The saved costs for the potable water were almost the same as the installation cost, meaning an instant return on investment [26] (Figs. 8 and 9).

Green roof research suggests retention values of NGRs are better than those quoted in the FLL

guidelines. One reason for this is that overload and vegetation of green roofs were not taken into account. Consequently, more research is required here. Bearing that in mind, it should seem obvious that NGRs do have a significant impact on storm water runoff reduction and mitigation. Green roofs are an important part of “green technology.” Yet green roofs should be part of a far larger sustainable development plan for the world’s cities.

The design and implementation of green roofs specifically focusing on the aspects of rainwater management are a new issue. Green roofs should be standard elements in the calculation of water management in buildings. The data from FLL [34] and other research should be basic elements in programs to model the ecological footprint of buildings. Green roofs must be integrated into policies, regulations, building codes, and other tools for sustainability.

To encourage further understanding of the benefits of green infrastructure on urban water systems, further studies will assist in indentifying regional idiosyncrasies. Green roofs work suitably in temperate and tropic climates, but local research facilities are helpful in identifying site-specific plants and soils. Green roof technology is an interdisciplinary technology for urban water management issues – more research with a focus on civil engineering would be further supportive, like Knoll [42]. Though green roofs do offer long-term benefits, minimum maintenance is required.

A before/after impact mathematical model of an NGR as a retrofit can be seen at, e.g., <http://www.sieker.de/modules/wfchannel/index.php?pagenum=5>

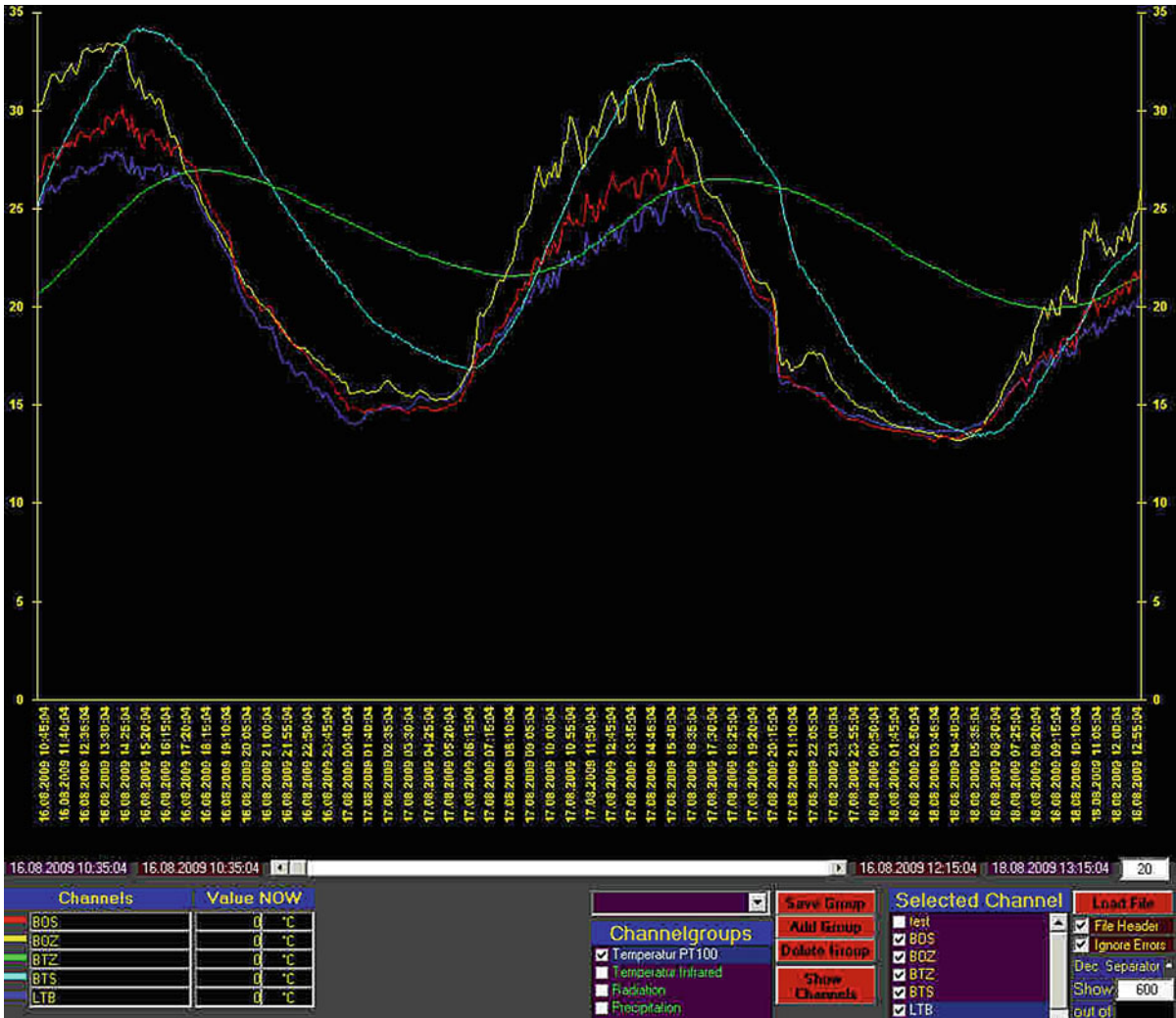
In the late 1970s, the German green roof industry’s position vis-a-vis the irrigation of extensive green roofs was that green roofs should not be irrigated.

In the 1980s, this position changed as rain water–harvesting concepts and technologies became popular in Germany. Today, green roofs are an integrated technology in the overall rainwater management concept. Evapotranspiration on an NGR, which is irrigated during the summer in a hot country and temperate region alike using water harvested from rainwater provides significant cooling effects in combination with rainwater management. Green roofs can provide an integrated solution to many of the challenges found in the modern built environment. In addition, summer irrigation allows for

greater diversity of plant species on a green roof. Yet, when the third factor of biodiversity is added to tailored solutions, it may be desirable to reduce irrigation (see “Biodiversity and Green Roofs”). What is becoming apparent is that green roofs can provide highly sophisticated solutions to many seemingly intractable problems faced by cities and green roofs are infinitely fine tunable to achieve very specific aims.

Energy Performance of Green Roofs

The thermodynamics and thermal performance of green roofs have been studied extensively since the 1970s. The Fig. 3 explains the typical temperature values of a surface and subsurface temperatures of a gravel roof and a dry extensive green roof on three selected summer days. Remarkable is the attenuation value of the subsurface temperature of the green



Green Roofs, Ecological Functions. Figure 3

Real summer surface and subsurface temperature performance over 3 days, 16th to 18th of August 2009 (Example of a screenshot of the measurements in Neubrandenburg x-axis: Temperature in 5° steps, from 0°C to 35°C Temperatures: Green line: temperature under the substrate Dark blue line: Air temperature at 1 m Light blue: underneath Gravel yellow: Surface gravel Red: surface Green roof

roof (green line, “BZT”, Fig. 3). The surface temperatures of black bitumen roofs have been compared with green roofs structures. A bitumen roof surface can exceed 70°C [43]. The Research roof in Neubrandenburg, see Fig. 3, compares extensive dry green roofs with a gravel layer on non-greened roofs. This facility has a data set from a longitudinal study of about 10 years with thermal values at 5 min intervals. The graph showcases three extremely hot summer days without rain. The green line is the surface temperature underneath the growing media. The red line shows the surface temperature of the green roof. The gravel temperature is shown by the yellow line. It can be seen that the red and yellow lines follow the air temperature. The air temperature, measured in a ventilated shaded shelter, has the same peaks and troughs but at a lower level. If the growing media was damp, these graphs would be almost identical. The light blue (gravel temperature) and the green line (growing media temperature) show the impact of green roofs clearly.

Roof surface building materials are deteriorated and age through exposure to ultra violet and heat. Consequently, it is obvious that a green roof necessarily extends the life expectancy of these materials.

Furthermore, the thermal performance of green roofs in terms of summer cooling and winter insulation, which reduces building energy requirements will obviously become more and more important as the depletion of finite energy resources continues. Germany has instituted strict energy-saving policy in recent decades, and green roofs have spearheaded this drive. [43] calculated the additional insulation effect of a 10-cm layer of growing media. The result was a reduction in energy consumption during winter months in Central Europe, equivalent to that achieved by an extra 1 cm conventional insulation. The thermal impact of green roofs depends on the building being studied. Parameters such as roof/wall juxtaposition are important. A detached house with standard insulation achieves an energy saving equivalent to a “week of free heating.” In light of increasing energy prices, this is an important additional benefit [43]. Also in Vancouver, a total heat flow reduction up to 70% was measured for 30 days in a moderate spring [44].

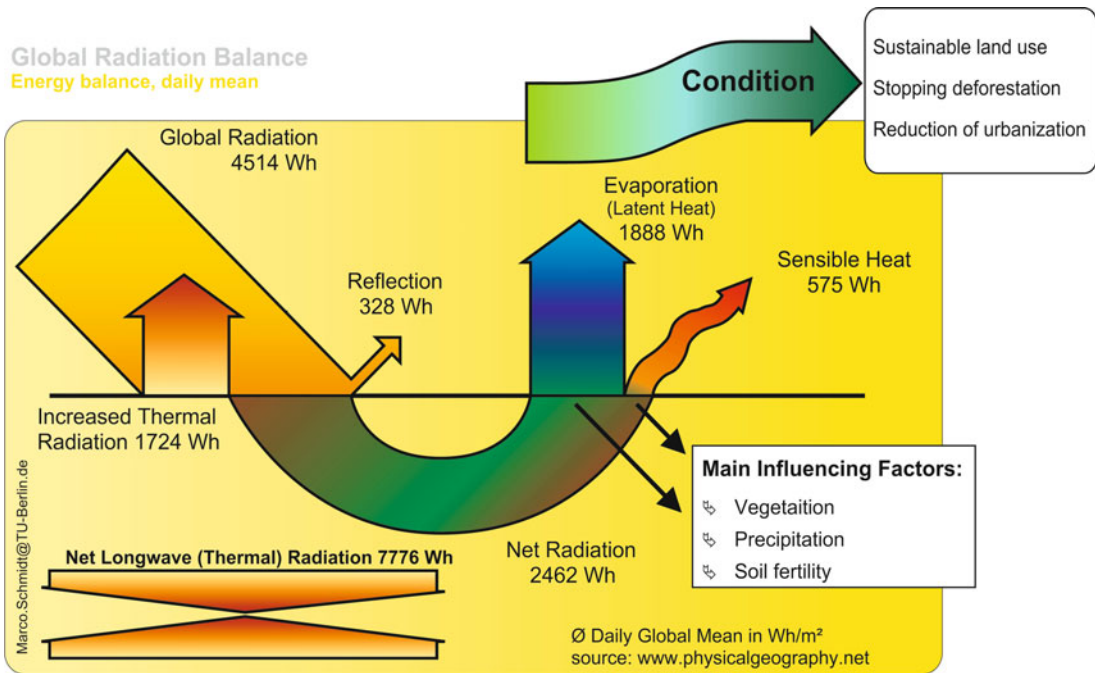
Green roofs really come into their own in the summer in reduced air conditioning requirements. Air conditioning use in Europe is on the rise, so this benefit

of green roofs could reverse this trend. The “*Report on Energy Efficiency and Certification of Central Air Conditioners*” [45] of the European commission is a good database concerning the development of the air conditioning market in Europe. In recent years, energy consumption for heating has decreased due to the use of insulation and other energy efficiency strategies. During the same period, the use of air conditioning increased at a rate of 12% per year in Germany. In Germany, an increase in energy consumption for cooling and ventilation of 260% is estimated until the year 2020 [31]. In stark contrast to this, the “climate protection program” of the German government set the target for fossil fuel use reduction at 40%. This ambitious goal is mainly based on the reduction of energy consumption in the building sector.

The poor energy efficiency of air conditioners driven by electricity further contributes to the problem. Cooling a room through electricity releases an even higher amount of energy outside of the building. This approach to cooling contributes to a further increase to the urban heat island effect. However, such an approach is not necessary, because an alternative exists.

On a global scale, evaporation of water is the largest and most important component for the conversion of solar radiation. It is also the largest hydrological component together with precipitation. Only water that evaporates causes rainfall. The evaporation creates a large and small water cycle of condensation and precipitation [31].

Figure 4 shows how global radiation components are converted on the surface of the earth. The data shown represents one square meter. The diagram shows that 328 Wh (7.3%) is reflected, and 1,724 Wh (38%) is directly converted to thermal radiation due to the increase of surface temperatures. The total long-wave or so-called thermal radiation consists of the atmospheric counter-radiation and the thermal radiation of the surface. The net radiation can either be converted into sensible heat or consumed by evaporation, which is the conversion into latent heat. At 1,888 Wh/(m²d), the energy conversion by evaporation is most important, even more important than the thermal radiation. Additionally, the evaporation influences the thermal radiation due to the change in surface temperatures. The atmospheric counter-radiation of 7,776 Wh is a theoretical component and can be



Green Roofs, Ecological Functions. Figure 4

Global daily radiation balance as annual mean [31]

considered as an exchange of long-wave radiation between two surfaces.

Heat flux data were taken at the study green roof in Neubrandenburg. Similar research was conducted by Karen Liu in Canada [44]. Both of these studies show that the heat flux in the summer is greater than in the winter. Consequently, the cooling effect of green roofs in the summer is even more important than the insulation effect in the winter.

When green roofs are irrigated, the evaporation effect results in major cooling, making green roofs natural cooling systems (Figs. 4 and 5).

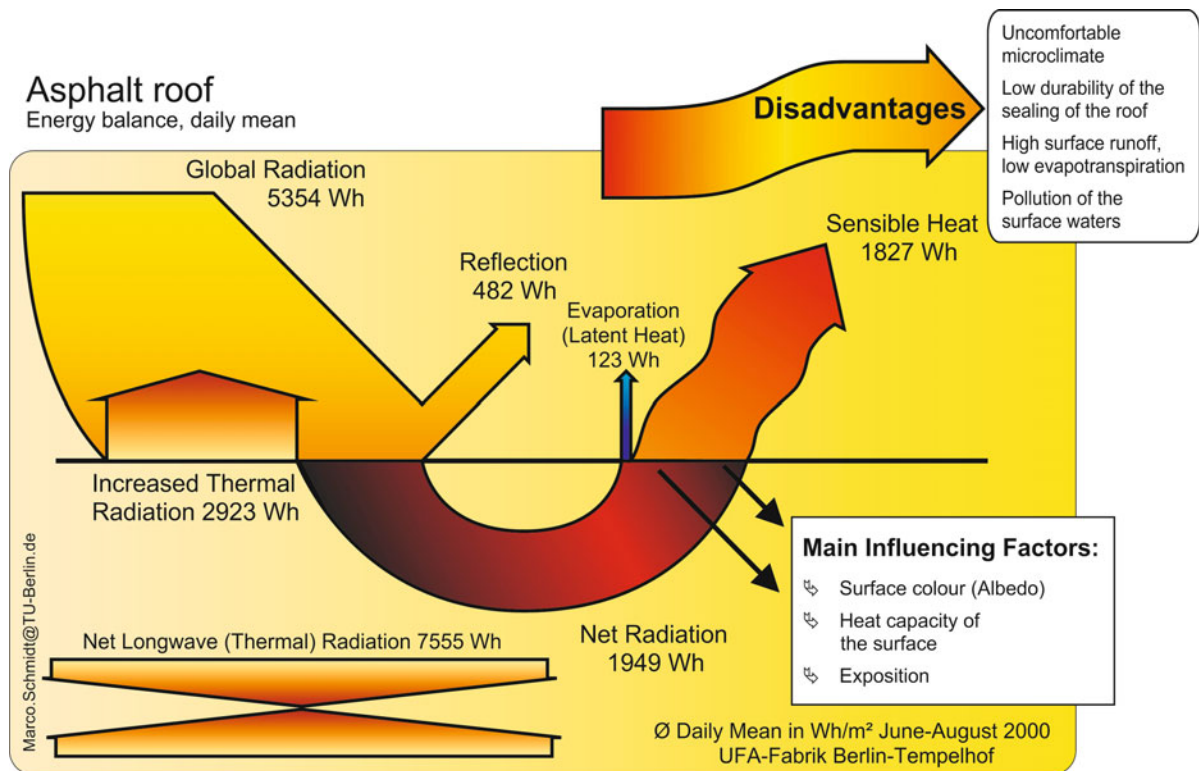
Conventional non-vegetated roofs have less evapotranspiration in cities. This is the reason for more direct sensible heat causing hotter surfaces at the same time. The deforestation that is occurring worldwide means losses in evaporation and higher temperatures. The two most promising ways of reversing global warming trends are urban greening and reforestation. On the global scale, the reduction in evaporation is what is mainly responsible for climate change [29 (Kravčik et al. 2007; www.waterparadigm.org)]. There is a rise in global temperatures in conjunction with increased

CO₂ emissions. Both of these states are caused by reduced vegetation on the Earth and increased human development [46]. The correlation between CO₂ levels and the global temperatures represents the relation between the amount of biomass, i.e., vegetation, the photosynthesis process, and the evapotranspiration of that biomass.

The evaporation of water is the cheapest and most effective way to cool surfaces. If the rate of evaporation can be slowed, the effect is more efficient and longer lasting. This is what a green roof does. Vegetated structures evaporate water and use CO₂. Trees are the most effective cooling systems [47]. Trees must be planted in cities on a massive scale in every available space. Massive implementation of living wall and green roof technology must also be implemented.

The evaporation of one cubic meter of water requires 680 kWh of heat. Green roofs are the most efficient way to cool down a city.

Rainwater harvesting techniques, which focus on evaporation rather than storage could play a key role in further adaptation and mitigation strategies against the urban heat island effect and global warming. Green



Green Roofs, Ecological Functions. Figure 5

Radiation balance of a black asphalt roof as an example for urban radiation changes [31]



Green Roofs, Ecological Functions. Figure 6

Net Radio meter. Details from the research measurement facilities on a green roof at the University of Applied Sciences, Neubrandenburg, one of the Net – radiometers installed above different roof types to calculate the radiation balance

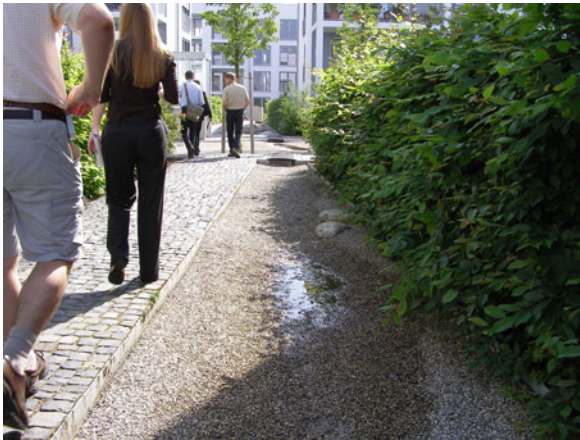


Green Roofs, Ecological Functions. Figure 7

Lysimeter. Detail of the roof lysimeter installation to measure the real water content of roof growing media and gravel roof



Green Roofs, Ecological Functions. Figure 8
Block 6. Plant sewer system in Berlin with rainwater cleaning facility in the back yard of a residential area of the City Center of Berlin, Details on this project can be seen at: http://www.stadtentwicklung.berlin.de/bauen/oekologisches_bauen/download/modellvorhaben/flyer_block6_engl.pdf



Green Roofs, Ecological Functions. Figure 9
Abb. 4. An example of a rain garden infrastructure on an intensive Green roof/Roof garden in Stuttgart

roofs and green facades have a huge potential for decreasing the environmental impact of urbanization. On a global scale, the reduction in evaporation is main cause of climate change [(see www.waterparadigm.org)]. Simulations of global climate changes still neglect the main driving factor for the global climate: the evapotranspiration of vegetation. Using

evaporation, cooling for cooling purposes on green roofs and redirecting the cooling load into the dwellings mean a 41–93 higher efficiency of this energetic process than producing environmental cooling by electricity [48, 49].

Biodiversity and Green Roofs

The conditions on a roof are characterized by higher wind speed, and high direct sun irradiation. To establish vegetation on roofs, it is helpful to have higher parapet walls to protect the plants to establish the plant cover more easily. In addition, pergolas and other supporting structures help the plants grow better.

Roof gardens provide the opportunity to select shrubs and small trees. Roof gardens must be carefully designed with the knowledge of the mature heights and sizes of selected plants along with their maintenance requirements. Plants with aggressive root systems should be avoided. Bamboo is a popular plant, nowadays, which is a disaster on roof gardens because of its aggressive root system, invasive nature, and uncontrollable growth.

Roof gardens or intensive green roofs are known as “urban forestry” in Asia. A wide range of plant species are suitable for roof gardens, which fulfill the design conditions mentioned.

On the other hand, ground cover vegetation species variety is more limited. Research by Tan [49] suggested that succulent plants from South Africa are suitable for green roofs throughout the hot humid tropics. Well-selected dry adapted plant species can survive for about 49 days in the Singapore without water input from precipitation or irrigation, and some plants can go beyond this and although they look dead will recover when water comes [50].

Natural green roofs (EGR) are extreme ecosystems [51]. Many NGRs have low static loads with a shallow layer of artificial growing media of about 5–10 cm in depth. After installation, they are often nearly maintenance-free. Normally, a little irrigation is used in the first year to support and encourage the initial vegetation cover. In subsequent years, only a single annual inspection is necessary to identify any maintenance needs. These needs involve mainly the removal of tree seedlings and erosion control. The vegetation will adapt

to the local climate situation after a short while. It will grow to a nearly full vegetation cover without any additional irrigation and fertilizer for a long time.

Green roofs are completely man made. The negative impact of this urban ecosystem is one of the lowest of all built-up areas in cities.

About 30 years ago, decentralized oven heating emitted so much pollution, which included micro nutrients, such as sulfur and several others, that the EGRs had a good nutrition supply. When “Clean Air Policies” were instituted in cities, a nutrient deficit occurred, and a debate began within the green roof guideline commission in Germany [34] about the need for fertilizing EGRs. Poor nutrient availability is good for extreme plant species. However, added fertilizers could help to achieve a more complete vegetation cover.

The nearly zero maintenance of typical EGRs is good for wildlife as well for establishing plants, which are adapted to poor and extreme climatic conditions found on roofs with higher radiation and higher wind speed. The names for EGRs are different all over Europe, and include the following ‘living roofs’ ‘oikosteges’ “Eco roofs,” “Sod roofs,” or “Grass roofs.” In Northern European NGRs, the percentage of grasses is higher; the modern NGR in the last 20 years is dominated by *Sedum* species in Germany.

Only a small number of plant species are suitable for extreme NGRs without any irrigation. Most of the adopted plant species have their origins on sandy dunes or poor soils and wall structures in marginal agricultural land [52]. In normal cases, it is a mixture of mosses and lichens combined with a small number of vascular plants. Succulents are one of the most well suited NGR plants. Consequently, many designers, particularly in Germany and German-influenced regions, use *Sedum* species. Specialized nurseries offer a wide selection of *Sedum* for green roofs [53,54]. But also a number of grasses can dominate such roofs, where the water storage capacity of the growing media becomes important.

The vegetation of the early period of modern green roofs constructed before the World War II was investigated by Kreh et al. [55–57]. These researchers came to the conclusion that a substrate layer of about 10 cm is suitable for a *Poetum compressa* – meadow as the climax vegetation.

Re-building and renovation work can cause damage to the vegetation and what is observed is the arrival of annual plants as biological succession begins again. Urban birds such as pigeons and doves, bring in seed of several pioneers, which will grow there for a while on the roof. The work of Darius and Drepper [57] correlated soil depth in centimeters and the successful establishment of vascular plants. This research concluded that at a depth of less than 10 cm, there will be a cover of mosses and *Sedum* in Germany. Grasses began to become established when substrate depths were 10 cm, and deeper than 20 cm substrate depths can support brown-field-grasses like *Calamagrostis epigejos*. Tree seedlings are in all cases possible, but they can only survive for extended periods if the roots can grow into cracks of the walls for a better water supply in dry seasons. So trees on the roof are, in most cases, a green roof problem as they compromise the waterproofing.

There are some 100 year old + EGRs [58]. Many of these will be removed in the coming years as these buildings are renovated. The potential value of these old green roofs is un-estimable as they are living examples of how a green roof develops and so should be protected in the same way as other natural ecosystems are protected. Research interest in such old German natural green roofs comes from outside Berlin.

At the University of Sheffield irrigation, research was conducted [59] and concluded that plant selection and depth of substrate were the vital characteristics of rich biodiversity and successful natural green roofs in the British climate. The most well suited species the study concluded were a number of *Sedum* species, *Festuca*, and species such as *Armeria maritima* and *Prunella vulgaris*.

In the more arid climate of Eastern Germany, the last two plant species are not able to survive on an average extensive green roofs. A 20-year longitudinal study concluded that in Berlin, for example, Chive (*Allium schoenoprasum*) was one of the most suitable species [51] (Tables 3 and 4).

To achieve a high value in local biodiversity, some aspects are important, like:

- Location (where is this project located)
- Size of the greened areas of the project

Green Roofs, Ecological Functions. Table 3 Plant species selection for EGR

Endangered Plants, growing well on NGRs (temperate climate of Northern Europe)	<i>Saxifraga tridactylitis</i> [55 Kreh]
	<i>Teesdalia nudicaulis</i>
	<i>Poa bulbosa</i>
Some of the fittest in Central Europe	<i>Sedum album</i> and many more of the genus <i>Sedum</i>
	Cloves, many Fabaceae are good competitors on poor and dry soils
	Grasses, like genus <i>Festuca</i> , <i>Poa</i> ,
	<i>Allium schoenoprasum</i>
Fittest plant species in subtropics and tropics, selected after [49, 50]	<i>Delosperma cooperi</i> , <i>Cyanotis cristata</i> , <i>Wedelia trilobata</i> , <i>Liriope muscari</i> , <i>Ophiopogon japonicus</i> (This is a favorite in Germany for indoor living walls, Koehler).
	<i>Bryophyllum</i> "Mother of Thousands" – Crassulaceae

Green Roofs, Ecological Functions. Table 4 Groups of plants, considered to be troublesome weeds, which should be removed during the annual maintenance work

		Recommendation
Tree seedlings	<i>Birch</i> , <i>Prunus</i> , <i>Salix</i> , <i>Hippophae</i>	Annual inspection
Grasses	<i>Agropyron</i> , <i>Calamagrostis</i>	
Herbs	<i>Taraxacum</i> ,	
Mosses	Several taxa, removal needed, must be discussed in relation to the project aims.	

- Variation of growing media, and microhabitat
- Plant species sources in the surroundings of the project
- Age of the project, young projects include for the first years a high number of annual plant species, mature projects could have well-established vegetation

Ecological Synergy

A total overview of all ecological benefits of green roofs cannot be described for all regions in the same way. Table 5 summarizes some conclusions from several research projects.

Noise reduction, green roofs are also a good sound proofing technology and for this reason, they are used when there is a lot of noise like Schipol airport in Amsterdam.

Biodiversity: On the research roofs of the University of Neubrandenburg, the plant species richness has been investigated since the installation of the green roof structure in the year 2001. The experimental variables are four orientations (North, South, West, and East), two growing media, two methods of planting (seeds and pre produced lawn mats). The results of this study suggest the following: There are significant differences in vegetation development between the variable "North" and "South." On North, there are more grasses and some interesting herbs. On the South experimental roof, different species of the genus *Sedum* dominate. Grasses are reduced to a very few individuals.

Furthermore, over a period of the first 7 years, a significant difference between the turf mat installation and the *Sedum*-seeded area was observed.

The seeded roof developed higher species richness. The species richness of the preproduced turf mats decreases over this time [65].

Green roofs have become a topic of great research interest by a plethora of disciplines. There are about 500 peer-reviewed publications as of 2010 in various disciplines (a M. Kohler survey with Springerlink and similar data bank systems).

One interesting recent study focused on the topics: Do green roofs affect the quality of the water discharged during a storm event? Can this water be used? [66]. The answer to this question is that it depends on a number of factors. One important conclusion is that the additives used by the construction industry will become scrutinized more carefully in the coming years. Green roofs can remove numerous heavy metals from rainwater.

- The cooling effect of green roof systems must be studied extensively in the various regions of the World. Fang [67] worked on the conditions of Taiwan with climate chamber experiments.

Green Roofs, Ecological Functions. Table 5 Main ecological effects of Green roofs [60, updated]

			Monetary efficiency	Source	Private/ public good
1	Insulation, Winter, Central Europe,	3–10% better than a gravel roof, 100 m ³ gas reduction each year for a single house	40.0 €/Building	Köhler (2009) [60]	Private
	Winter, southern Europe,	Depending on the type of building, no effect	–	[61] Spala	Private
2	Air can substitute summer	Insulation effect and evaporation cooling of about 3 l/m ²	Up to 20% reduction, depending on the climate in upper level of a multistorey house	[62] Alcazar	
3	Urban climate	Cooler surfaces in summer, better insulation	Monetization, done by Banting et al. in Toronto. A green roof calculator exists on www.greenroof.org under development	[63] Banting	
4	Noise reduction	Test facilities are installed in Vancouver	A reduction of about 5 dB(A), depending on structure, moisture, and some more	[64] Maureen	
5	Biodiversity	Biodiversity depends on various factors	No monetary effect	[65] Kö Baltimore	

The conclusion of this study stated that a thermal reduction of surface temperatures between 20% and 60% was possible depending on the depth of the substrate, its moisture content, and the vegetation type and amount of coverage.

- The thermodynamic performance of NGRs in Greece is now being studied by the Metsovio National Technical University of Athens. They have studied the oikostegi installed on The Greek Treasury in Constitution Square in Athens, opposite the Greek Parliament. They concluded among other things that the oikostegi has significantly altered the thermodynamics of the 10-story building. http://www.oikosteges.gr/index.php/green_roofs/research.
- Binding fine particles is another benefit of green roofs, which is currently being studied. Yang [68] conducted research in Chicago. He concluded that an area of about 20 ha of green roofs can capture about 1,675 kg/m² 10 particles each year. Gas born emissions, like ozone and NO_x, can be sequestered at of about 27% of total emissions.

Research interest in the subject is growing around the world. Moreover, mathematical and computer

modeling algorithms are also improving leading to sound data and conclusions, which can be used to support this exciting emerging technology.

Future Directions

At Climate conferences, like the last one in Copenhagen in November 2009, participants are searching for solutions regarding how to mitigate climate change. There will not be only one strategy. But it is a truth that global deforestation and the ever increasing levels of urbanized land all over the world need strategies to increase the amount of vegetation wherever it is possible. The quality of an “urban forest” in street canyons and on the surface of buildings is not the same. From the perspective of biodiversity, only a selected number of species are able to survive on shallow substrates. So weight must be give to endemic species.

The increasing number of urban surfaces offers opportunities for urban gardening and to establish specific vegetation, in some cases, also for some endangered plant species. Roof vegetation is an engineered system, which offers benefits, as described in this entry. Better insulation against cool and hot temperatures, rainwater management, and longer lasting roof

surfaces all provide monetary benefits for the property owner and all citizens. Finding the best technical solutions requires environmental specialists who understand natural ecosystems in conjunction with construction experts. As this technology grows, developments will be made, which will further improve the benefits [18]. Green roofs organizations and associations are now found in more than 30 countries worldwide. In many of these countries, academics are doing research into green roofs. Green solutions have become the hot subject of the twenty-first century. This is certain to continue as climate change begins to affect our everyday lives. Green roofs offer a way forward, which gives hope that sustainable technology is not only desirable but also very feasible. The further direction is to combine sustainable technologies and materials to enhance the number of plant species and ecological cycles on green roofs. The technical term “Natural Green Roof” will challenge urban planners in coming years. Green roofs are a fixed main and integrative element of sustainable architecture. It is cross-disciplinary subject, which is both simple and at the same time high tech. Much more growing media and plant species are possible. Also many more architectural solutions are possible [69]. It is also a two-way learning opportunity for an exchange between industrial countries and developing countries alike. In this direction, green roofs play an integrative role in rain water savings and management [70]. Green roofs provide a protection layer against temperature extremes, which provide energy-saving opportunities worldwide.

Acknowledgements

Many thanks to Marco Schmidt for his ongoing support and as a joint researcher in much of the work Manfred has done over the years.

Bibliography

Primary Literature

- Alexandri E, Jones P (2006) Temperature decreases in an urban canyon due to green walls and green roofs in diverse climates. *Build Environ* 41(4):480–493
- Niachou A, Papakonstantinou K, Santamaouris M, Tsangrassoulis A, Mihalakou G (2001) Analysis of the green roof thermal properties and investigation of its energy performance. *Energy Build* 33:719–729
- Theodosiou TG (2003) Summer period analysis of the performance of a planted roof as a passive cooling technique. *Energy Buildings* 35:909–917
- Oberndorfer E, Lundholm J, Brass B, Coffmann R, Doshi H, Dunnett N, Gaffin S, Köhler M, Liu K, Rowe B (2007) Green roofs as urban ecosystems: ecological structures, functions, and services. *Bioscience* 57(10):823–833, www.biosciencemag.org
- Köhler M (2005) The green roof movement – from a botanical idea to a new sustainable style in modern architecture. *Proceedings of 1th international landscape education symposium*, Shanghai, China, p 168–176. (ISBN 7-112-08650-7) www.china-building.com.cn
- Kellert SR, Wilson EO (1993) *The biophilia hypothesis*. Island Press, Washington
- Larson D, Matthes U, Kelly PE, Lundholm J, Gerrath J (2004) *The urban cliff revolution*. Fitzhenry and Whiteside, Markham
- Odum HT (1994) *Ecological and general systems. An introduction to systems ecology*. University Press of Colorado, Niwot
- Köhler M, Barth G, Brandwein T, Gast D, Joger HG, Seitz U, Vowinkel K (1993) *Fassaden- und Dachbegrünung*. Ulmer, Stuttgart
- Getter K, Rowe B (2009) Carbon sequestration potential of extensive Green roofs. *Proceedings of greening rooftops for sustainable communities*. Atlanta
- Köhler M (2009) Der Gründachmarkt weltweit. *Tagungsband 7. Internationales FBB Gründachsymposium 2009*, Ditzingen, pp 37–40
- Yeang K (2008) *Ecodesign – a manual for ecological design*. Wiley-Academy, Hoboken
- Kellert SR (2005) *Building for life*. Island Press, Washington
- Todd NJ, Todd J (1993) *From eco-cities to living machines*. North Atlantic books, Berkely
- Peck S (2008) *Green roof designs*. A. Schiffer book, Atglen (PA- USA), p 176
- Weiler SK, Scholz-Barth K (2009) *Green roof systems*. Wiley, Hoboken
- Jim CY, Chen WY (2009) External effects of neighbourhood parks and landscape elements on high-rise residential value. *Land Use Policy* (Elsevier Science, Amsterdam) 27:662–670
- Jodidio P (2009) *Green architecture now*. Taschen, Hongkong, p 416
- Koehler M, Schmidt M, Grimme FW, Laar M, De Assuncao Paiva VL, Tavares S (2002) Green roofs in temperate climates and in the hot-humid Tropics. *Environ Health* 13(4):382–391
- Ansel W (2008) A tale of 3 cities – comparative analysis of green roof policies and success factors, Cuge regional seminar. Oct 23th, Singapore
- Appl R, Meier R, Ansel W (2009) *Dachbegrünung in der modernen Architektur*. *Proceedings of IGRA*, Berlin
- Göbel P, Dierkes C, Kories H, Messner J, Meissner E, Coldewey WG (2007) Einfluss von Gründächern und Regenwassernutzung auf Wasserhaushalt und Grundwasserstand in Siedlungen. *Grundwasser – Z. der Fachsektion Hydrogeologie* (12):189–200

23. Sukopp H, Wittig R (1997) *Stadtökologie*. Gustav Fischer Stuttgart, New York
24. Varis O, Biswass AK, Tortajada C, Lundquist J (2006) Mega cities and water management. *Water Resource Development* 22(2): 377–394
25. Reichmann B, Nolde E, Leithaus J, Vansbotter B (2002) Maßnahmenkatalog Reduzierung der Wasserkosten im öffentlichen Bereich. Berlin, Senatsverwaltung für Stadtentwicklung Berlin. http://www.stadtentwicklung.berlin.de/bauen/oekologisches_bauen/de/downloads/massnahmenkatalog_wasserkosten.pdf. Accessed 26 May 2008
26. Reichmann B, Nolde E, Rüden H, Vansbotter E (2007) Innovative water concepts. Service water utilisation in buildings 28 S. http://www.stadtentwicklung.berlin.de/bauen/oekologisches_bauen/de/downloads/betriebswasser_englisch2007.pdf. Accessed 26 May 2008
27. Köhler M (2008) Green facades – a view back and some visions. Urban ecosystems. www.springerlink
28. Centgraf S, Schmidt M (2005) Water management to save energy, a decentralized approach to an integrated sustainable urban development. Proceedings of Rio05, Brasil. Accessed 26 May 2008
29. Kravčík M, Pokorný J, Kohutiar J, Kováč M, Tóth E (2007) “Water for the recovery of the climate – a new water paradigm. Publisher Municipalia. <http://www.waterparadigm.org/>
30. Köhler M, Schmidt M (2008) London benefits for sustainable water management. World green roof technology. Proceedings of world green roof congress, London, 16–19 Sept
31. Schmidt M, Koehler M (2008) Energetic aspects of green roofs. World green roof technology. Proceedings of world green roof congress, London, 16–19 Sept
32. Connolly M, Liu K (2005) Green roof research in British Columbia – an overview. Proceedings of greening rooftops for sustainable communities, Washington
33. Palla A, Gnecco I, Lanza LG (2010) Hydrologic restoration in the urban environment using green roofs. *Water* 2, 1. www.mdpi.com/journal/water ...
34. FLL (ed) (2008) Richtlinie für die Planung, Ausführung und Pflege von Dachbegrünungen. Bonn, English version: Guidelines for the planning, construction and maintenance of green roofing – green roofing guideline, 2008 edition
35. Liesecke HJ (1998) Das Retentionsverhalten von Dachbegrünungen. *Stadt + Grün* 47:46–53
36. Köhler M, Keeley M (2005) The green roof tradition in Germany: the example of Berlin. In: Leslie Hoffmann, WMc Donough (eds) *Ecological design and construction*. Earthpledge, New York, pp 108–112
37. Keeley M (2007) Transatlantic exchange and sustainable Urban development: transferring stormwater policies and technologies from Europe to the United States. Ph.D., Technical University, Berlin, 259 p
38. Köhler M (2004) Energetic Effects of Green roofs on the urban climate near to the ground and to the building surfaces. Proceedings of international green roof conference, Nuertingen, IGRA, S.72–79
39. Köhler M, Schmidt M, Grimme FW, Laar M, De Assuncao Paiva VL, Tavares S (2002) Green roofs in temperate climates and in the hot-humid tropics. *Environ Health* 13(4):382–391. (UK) ISSN 0956-6163
40. Bustorf J (1999) Simulation of the precipitation/runoff – ratio of greened roofs. Master thesis, Technical University, Berlin, 108 p
41. Köhler M, Schmidt M (2002) Das Mikroklima extensiver Gründächer. In: *Jb. Dachbegrünung 2002*: 28–33. Thalacker, Braunschweig
42. Knoll S (2000) Das Abflußverhalten von extensiven Dachbegrünungen. Mitt. Nr. 136 TU-Berlin, Inst. für Wasserbau und Wasserwirtschaft, 115 S
43. Köhler M, Malorny W (2009) Wärmeschutz durch extensive Gründächer. In: Venzmer H (*Europäischer Sanierungskalender 2009*). Beuth, Berlin, pp 195–212
44. Liu K (2008) Sustainability matters. U.S. General services administration. See www.Gsa.gov/P100
45. EECAC (2003) Energy efficiency and certification of central air conditioners. REPORT for the DGTREN of the commission of the E.U, 2001, Volume 1, 52 p
46. Gerlich G, Tscheuschner RD (2007) Falsification of the atmospheric CO2 greenhouse effects within the frame of physics. 114 p. <http://arxiv.org/abs/0707.1161>; <http://arxiv.org/pdf/0707.1161v3>; <http://www.tsch.de>
47. Currie BA, Bass B (2008) Estimates of air pollution mitigation with green plants and green roofs using the UFORE model. *Urban Ecosystem* 11(4): 335–337, Springer
48. Mankiewicz PS, Spartos P, Dalski E (2009) Green roofs and local temperature: how green roofs partition water, energy, and costs in urban Energy – air conditioning budgets. Proceedings of greening rooftops for sustainable Communities, Atlanta
49. Tan PY (2009) Understanding the performance of plants on non irrigated Green Roofs in the Tropics using a Biomass yield approach. *Nature in Singapore*. <http://rmbn.nus.edu.sg/nis>
50. Tan PY, Sia A (2005) A selection of plants for green roofs in Singapore. CUGE Singapore, 117 p
51. Köhler M (2006) Long term vegetation research on two extensive green roofs in Berlin. *Urbanhabitats*, Brooklyn Bot. Garden (USA) 4(1):3–26. ISSN 1541-7115. http://www.urbanhabitats.org/v04n01/berlin_full.html
52. Darlington A (1981) *Ecology of walls*. Heinemann, London
53. Stephenson R (1994) *Sedum cultivated stonecrops*. Timber, Portland
54. Snodgrass EC, Snodgrass LL (2006) *Green roof plants*. Timber, Portland
55. Kreh W (1945) Die Pflanzenwelt unserer Kiesdächer. Jahresheft des Vereins für Vaterländische Naturkunde in Württemberg 97–101:199–207
56. Bornkamm R (1961) Vegetation und Vegetationsentwicklung auf Kiesdächern. *Vegetatio* 10:1–24
57. Darius F, Drepper J (1984) Rasendächer in West-Berlin. *Das Gartenamt* 33:309–315
58. Köhler M, Poll P (2010) Life time performance of selected old green roofs in comparison to extensive green roofs in Berlin. *Ecological Engineering* 36:722–729

59. Nagase A, Dunnet N (2010) Drought tolerance of different vegetation types in extensive green roofs: effects of watering and diversity. *Landscape and Urban Planning* 97:318–327
60. Köhler M (2006) Extensive Gründächer – Rechenbare Vorteile in der Eingriffsregelung. *Stadt und Grün* 9:40–44
61. Spala A, Bagiorgas HS, Assimakopoulos MN, Kalavrouziotis N, Matthopoulos D, Mihalakakou G (2008) On the green roof system. Selection, state of the art and energy potential investigation of a system installed in an office building in Athens, Greece. *Renewable Energy* 33:173–177
62. Alcazar S, Bass B (2005) Energy performance of green roofs in a multi Storey residential Building in Madrid. *Proceedings of 3rd Conference on greening roof tops*, Washington
63. Banting D, Doshi H, Li J, Missios P (2005) Report on the environmental benefits and costs of green roof technology for the city of Toronto. (kann auf der Seite der Stadt Toronto als pdf geladen werden. www.toronto.on.ca/greenroofs)
64. Connelly M, Hodgson M (2008) Sound transmission loss of green roofs. *Sixth Annual Greening rooftops for Sustainable Communities Conference*, Baltimore
65. Köhler M (2008) Extensive green roof biodiversity: the influence of growing media, exposition and the methods of establishing. *Proceedings of Baltimore green roof for healthy city conference*. ISSN 1916-4734, 16 p
66. Berndtsson JC, Bengtsson L, Jinno K (2009) Runoff water quality from intensive and extensive vegetated roofs. *Ecol Eng* 35:369–380
67. Fang CF (2008) Evaluating the thermal reduction effect of plant layers on rooftops. *Energy Build* 40:1048–1052
68. Yang J, Yu Q, Gong P (2008) Quantifying air pollution removal by green roofs in Chicago. *Atmos Environ* 42: 7266–7273
69. Bartoli B (2008) *Sustainable dalla A alla Z. Sistemi editoriali*. AS 25 Napoli
70. Franken M (2007) *Gestion de aguas*. Plural editores, La Paz

Books and Reviews

- Dunnet N, Kingsbury N (2008) *Planting green roofs and living walls*, 2nd edn. Timber, Portland
- Ernst W (2005) *Dachabdichtung Dachbegrünung*. IRB-Fraunhofer Gesellschaft Stuttgart
- Krupka B (1992) *Dachbegrünung*. Ulmer, Stuttgart

Recommended Additional Internet Links

- <http://www.cbd.int/cop9/>
- <http://oikosteges.gr/index.php/studies>
- <http://www.gebaeudekuehlung.de>
- <http://www.a.tu-berlin.de/GtE/forschung/Adlershof>
- http://www.stadtentwicklung.berlin.de/bauen/oekologisches_bauen
- http://www.stadtentwicklung.berlin.de/bauen/oekologisches_bauen/download/modellvorhaben/flyer_block6_engl.pdf

- <http://www.waterparadigm.org>
- <http://www.evapotranspiration.net>
- www.worldgreenroof.org
- www.livingroofs.org

Greenhouse Gas Emission Reduction by Waste-to-Energy

BERND BILITEWSKI, CHRISTOPH WÜNSCH
Institute of Waste Management and Contaminated Site Treatment, Technical University of Dresden, Pirna, Germany

Article Outline

Glossary
Definition of the Subject
Introduction: Importance of Waste Incineration
Release of Greenhouse Gas Emissions by Waste Incineration
Determination Methods for Determining the Fossil-based and Biogenic Carbon Content of MSW
Waste Composition
Fossil Carbon Content of MSW
Global Warming Factors of Different Waste Types in Germany
Reduction of GHG Emissions and Conservation of Fossil Fuels by Means of Energy Recovery from Wastes
Resource Conservation by Savings of Fossil Fuels
Greenhouse Gas Emissions by Landfilling
The Germany Example
Greenhouse Gas Emissions by Waste Incineration in Germany 2007
Reduction of Greenhouse Gas Emissions in Germany by Means of Waste Incineration in 2007
Future Directions
Bibliography

Glossary

Carbon dioxide (CO₂) Is a by-product of the combustion of fossil fuels or organic materials. Carbon dioxide is the principal greenhouse gas (GHG) in earth's atmosphere.

Carbon dioxide equivalents ($\text{CO}_{2,\text{eq}}$) Greenhouse gases influence the GHG effect in different degrees. Their contribution is calculated in volume (or mole) equivalents to carbon dioxide.

Combined heat and power generation (CHP) Is the simultaneous generation of both electricity and useful heat. Energy at a high temperature level is first converted to electricity and the remaining energy, at a low level, is used to produce heat (e.g., district heating).

Global warming factor (GWF) Expresses the amount of released $\text{CO}_{2,\text{eq}}$ for a combusted unit of fuel, in $\text{Mg CO}_{2,\text{eq}}/\text{Mg}$ of fuel; can be expressed as generated amount of electricity, in $\text{Mg CO}_{2,\text{eq}}/\text{MWh}_{\text{el}}$, or heat, in $\text{Mg CO}_{2,\text{eq}}/\text{MWh}_{\text{th}}$.

Greenhouse gas (GHG) Gases in the atmosphere that absorb and reemit infrared radiation; they cause the GHG effect that results in heating up of the atmosphere.

Lower heating value (LHV) Also known as net calorific value (CV) of a fuel, defined as the amount of heat released by the combustion of a unit mass of fuel; the LHV assumes that the water produced in combustion is in vapor state and that the latent heat of vaporization is not recovered.

Municipal solid waste (MSW) Predominantly household wastes (domestic wastes) and similar commercial and bulky wastes collected by a municipality. They are in a solid or semisolid form, sludge or liquids are excluded as well as industrial wastes.

Nitrous oxide (N_2O) Also known as laughing gas. Is a major greenhouse gas with a 298 times higher impact factor than carbon dioxide in a 100-year period.

Selective catalytic reduction (SCR) Is the reduction of nitrogen oxides in combustion gases to nitrogen, using a catalyst and anhydrous ammonia, aqueous ammonia, or urea to diatomic nitrogen and water.

Wet weight (ww) Substance in its original state, including combustibles, water, and ash.

Definition of the Subject

Waste management in general plays an important role in the GHG reduction targets of every country. Usually, the generated municipal solid wastes (MSW) are landfilled and in many cases these landfills are not

controlled and their emissions are not collected. Collection and treatment of the emissions of regulated landfills can reduce their GHG effect by up to 75%. The reduction of wastes going to landfill and the replacement by a mechanical, biological, and/or thermal treatment lowers the released GHG emissions significantly.

By means of these processes, the contribution of waste management to the GHG emissions of a nation can be reduced from 3.3% to 1.2% (e.g., in Germany [1]). In addition, the use of waste incineration in a modern waste management system produces electrical power and heat from the wastes. This results in a further decrease of GHG emissions.

Introduction: Importance of Waste Incineration

The principal method of disposing MSW worldwide is landfilling. After the initial wastes are covered by other wastes, atmospheric oxygen cannot reach them and under the prevailing anaerobic conditions, the biogenic carbon in MSW is converted to methane (CH_4) and CO_2 . If the landfill is not equipped for landfill gas collection, this CH_4 is emitted into the atmosphere. The global warming potential of CH_4 , relative to CO_2 and for a time horizon of 100 years, is between 21 and 25 [2].

In developed nations, more and more landfills are equipped with a landfill gas collection system. The collected biogas is then used to generate electricity or heat. This results in reducing GHG emissions, both by reducing the emission of methane and also by the use of fossil fuels substituted by the landfill gas.

Even though modern landfills are secured by ground and surface covers, during operation or if landfill covers are broken after closing the landfill, hazardous substances can be emitted in the environment. The major task of waste incineration plants recovering energy (waste to energy – WtE or energy from waste – EfW) is the environmentally friendly disposal of waste and the destruction of hazardous substances during the combustion process. Modern air pollution control systems clean the flue gas to emission levels below the EU standards. However, the production of the greenhouse gas (GHG) CO_2 , resulting from the oxidation of carbon in the MSW, is unavoidable. Also, a small amount of another greenhouse gas, nitrous oxides, is emitted; although this amount is small, as compared to the CO_2 emissions, it has to be assessed, because NO_x has a lifetime of

114 years in the atmosphere and its global warming potential is 298, in a time horizon of 100 years [2].

The thermal energy in the WTE flue gases is transferred to water and steam in a boiler and is used to generate electricity in a steam turbine or for district heating. Overall, the combined effect of WTE, that is, reducing landfill emissions and generating electricity, results in a GHG reduction of 1 t of CO₂ per ton of MSW combusted, for landfills that do not capture landfill gas, or half a ton of CO₂ for those that do (www.wte.org).

Release of Greenhouse Gas Emissions by Waste Incineration

A modern waste incineration plant consists of the combustion chamber where the waste is converted into ash and hot flue gas. The wastes are used as fuel, and an airflow provides the oxygen needed for combustion. The thermal energy of the combustion gases is transferred to steam in the boiler and superheater tubes, and the steam is used to generate electricity and also process steam and district heating. The cooled flue gas is treated in a flue gas purification system, and the clean gas is emitted to the atmosphere – necessary oxygen. A fossil fuel, for example, natural gas or fuel oil, is used only for start-up and shutdown operations. Water is used in the water steam cycle and, in some cases, in the flue gas purification system. The cleaned gas consists of nitrogen, CO₂, water, and oxygen. The residues include the incineration ash and the flue gas purification residues. Figure 1 shows the incoming and outgoing material and energy flows in a waste incineration plant.

For the calculation of GHG emissions, the waste fuels and any use of auxiliary energy are taken into account. Air, water, and auxiliaries do not produce significant GHG emissions. The clean gas containing the GHG and the delivered energy amounts, substituting energy produced and therefore their avoided GHG, need to be considered.

Determination Methods for Determining the Fossil-Based and Biogenic Carbon Content of MSW

In order to estimate the carbon footprint of waste incineration, it is necessary to know the fossil-based

carbon content of MSW. The carbon contained in the biogenic fraction of MSW and the corresponding CO₂ emissions are, by definition [4], classified as carbon neutral. The determination of the biogenic and fossil parts of the carbon can be made by means of a direct procedure (determination by the composition of the waste mixture) or an indirect one (determination in the flue gas) as discussed below.

In *manual sorting*, the waste mixture is separated to different waste fractions, and the biogenic and fossil parts in these single fractions are determined.

Selective extraction is a wet chemical procedure where oxidation of biogenic to non-biogenic components is used to identify the biogenic part of the waste mixture.

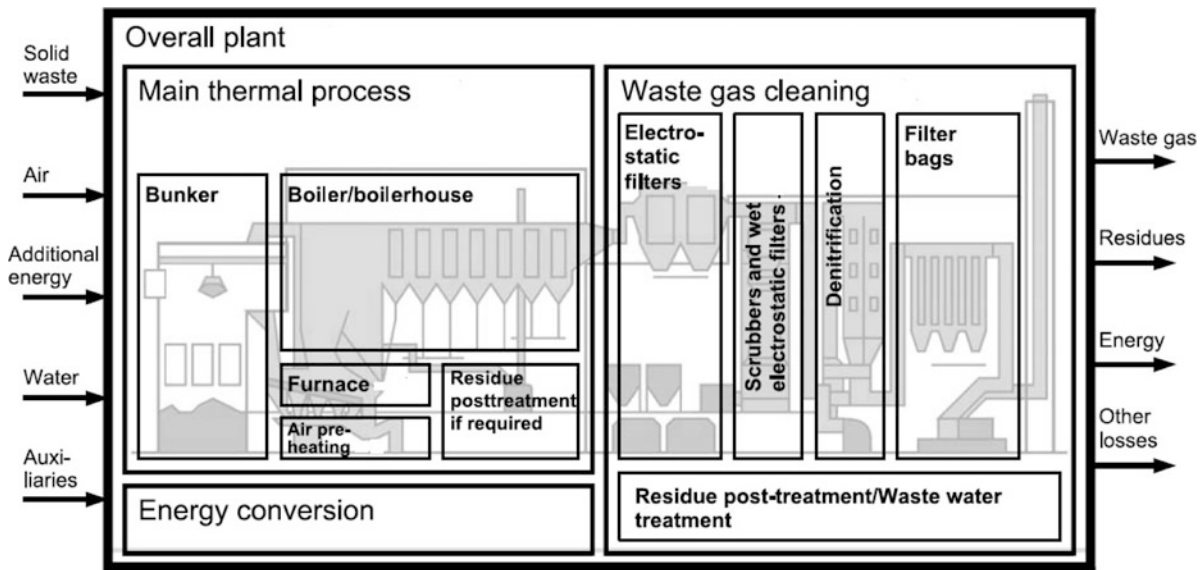
The ¹⁴C or *radio carbon method* is based on the difference in the age of fossil and biogenic materials. Because of the decay of the radioactive carbon isotope ¹⁴C, its concentration decreases with time. Because of its age, fossil carbon contains no ¹⁴C, whereas biogenic carbon contains a relatively higher ratio of ¹⁴C/¹²C. Therefore, the biogenic part in the CO₂ of the WTE stack gas is proportional to the ¹⁴C content, which can be measured by an instrumental method.

The *balance method* combines a theoretical mass balance with measurable operating data of an incineration plant. Every balance equation is characterized by a specific feature of the waste, for example, ash content, carbon content, etc., and the waste input is theoretically divided into four groups: inert, biogenic, fossil, and water. By solving these mass balance equations and with the help of nonlinear correction calculations, the biogenic carbon content can be determined.

The literature shows that their results are in the same range [5]. The manual sorting method is used in this section. The carbon isotope techniques are discussed in another section of the Encyclopedia.

Waste Composition

The basis of all calculations regarding released GHG emissions of wastes by the manual sorting method is the waste composition. First of all, the aggregate waste has to be separated in their different waste fractions. This is usually done manually. The composition is dependent on the kind of waste



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 1
Material and energy flows in a typical waste incineration plant [3]

(municipal solid waste, commercial waste, industrial waste, etc.) and, especially for the municipal solid waste (MSW), on the following:

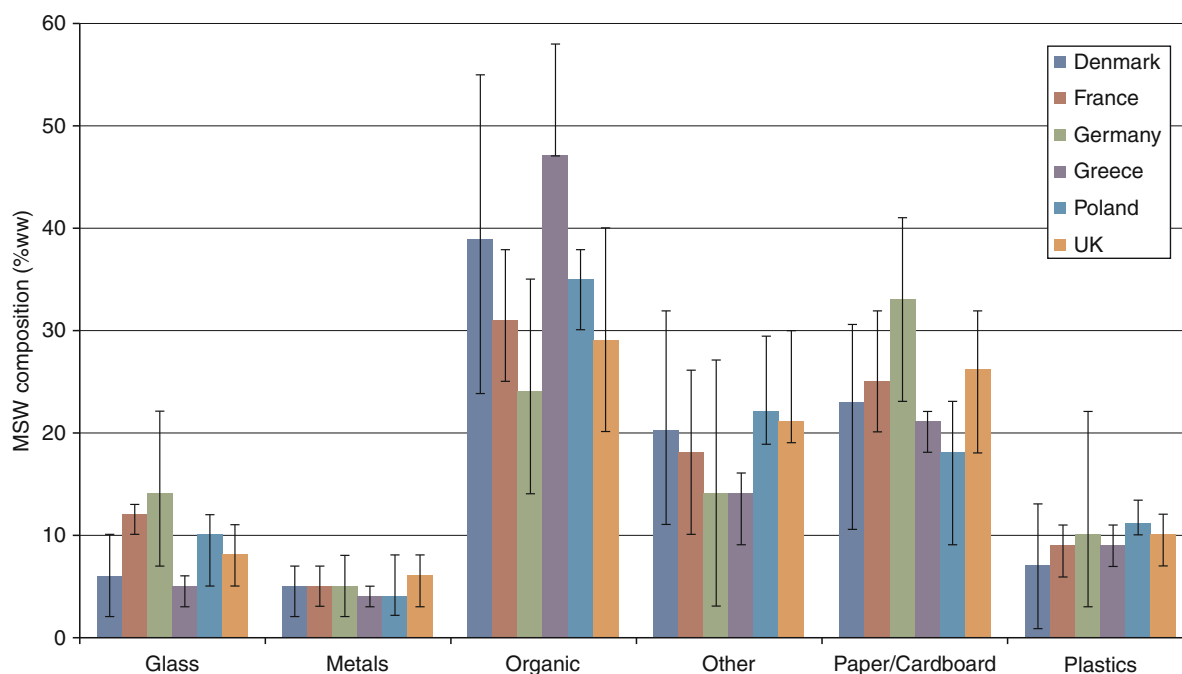
- Country, region
- Climate
- Settlement structure
- Season
- Economic situation
- Waste collection system

In different countries, rules and laws are existing influencing the waste management and finally the waste composition. As an example, in countries requiring a deposit on several materials like glass, plastic bottles, or cans, people source-separate these materials to get their money back. As a result, only a small part of such materials end up in the MSW. Also, when special bins for organic wastes, glass and paper are provided in a region or country, the amount of these materials in the MSW are reduced appreciably. In regions with hot and dry climate, the generated MSW contains a smaller fraction of organics than in wet climate areas. The reason is that fruits and vegetables grow well in these wet areas and after consumption the residues end up in the MSW.

Organics are often composted, and the compost is used in the gardens of rural areas; thus, this fraction is very small in the MSW. In urban areas, where there is no possibility of composting, the organics remain in the MSW. Another example is the higher amount of minerals and ash in the MSW of communities where coal or wood is used for heating. Also, during the winter season the amount of organic garden waste decreases while in the summer and autumn this fraction increases significantly. Ash residues also disappear in the summer months.

Economic development also results in greater consumption and generation of wastes, such as plastics, cardboard, and other packaging materials. Modern waste collection systems with single collection of glass, paper, cardboard, and plastics, as well as the possibility of sending back the used products to the manufacturers also reduce the amount of MSW and also influence their composition.

MSW is very heterogenic in nature, and no standard methodology exists for defining the waste composition [6]. It is therefore difficult to compare compositions resulting from different types of sorting analysis. Information about MSW composition in different EU countries was compiled by Gentil [7] and is shown in Fig. 2.



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 2

MSW composition in various European nations (Gentil, [7])

As illustrated in Fig. 2, the organic and paper/cardboard fractions constitute about 50% of the wet MSW, as received. Glass and plastics amount to about 10% each, and metals about 5%. “Other” materials (minerals, pollutants, composite materials, textiles, and fines) range from 15% to 20%. The error bars show the minimal and maximum values observed in a particular nation due to seasonal and other conditions.

Fossil Carbon Content of MSW

The total amount of carbon (TC) in MSW is estimated on the basis of the CO₂ released in combustion tests. The biogenic and fossil carbon parts can be estimated from the overall composition of the MSW or by carbon-14/carbon-12 measurements of the WTE stack gas, as described in another section of this Encyclopedia. The only materials containing fossil carbon are composite materials, plastics, and some textiles. Figure 3 shows the measured fraction of fossil carbon in various materials contained in MSW, as reported in the literature [7–13].

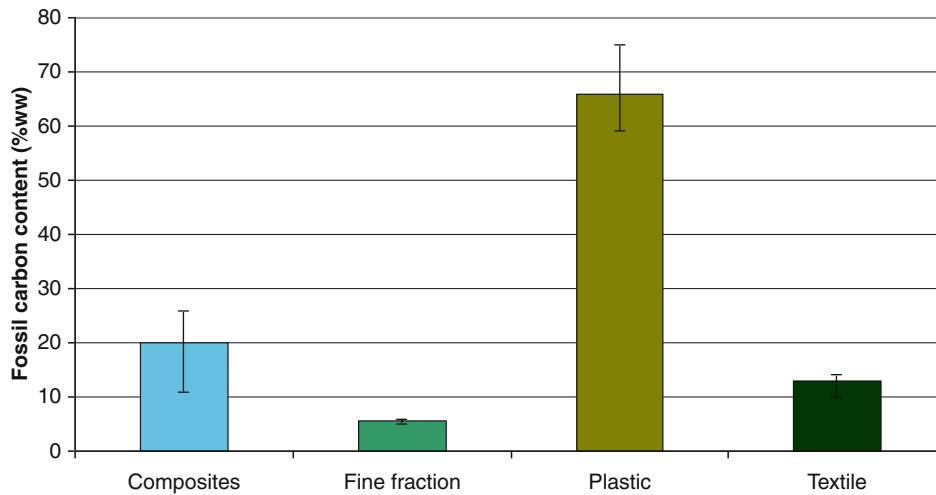
Due to the polymeric structure and the low water content of plastic, the concentration of fossil carbon is

very high (65–75%). The composite, textile, and fine fractions contain both biogenic and fossil-fixed carbon, and therefore their fossil carbon content is relatively low.

Global Warming Factors of Different Waste Types in Germany

The MSW collected by a municipality consists mostly of household wastes, but also includes bulky wastes, commercial waste that is similar to household wastes. In many communities, there is source-separation and collection of recyclables that are transported to materials recovery facilities (MRF) where they are separated either mechanically or manually to different recyclable streams. Between 2005 and 2007, the Institute of Waste Management and Contaminated Site Treatment of the Technical University of Dresden and the Intecus GmbH carried out an intensive campaign to analyze the waste types of household waste, similar commercial waste, and bulky waste all over Germany. The results of this investigation are shown in Table 1 [10].

Compared to the data that were shown in Fig. 2, Table 1 includes more waste fractions but the values are in the same range. The advantage of more detailed



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 3

Fossil carbon content of different materials contained in MSW

analysis is in providing a more accurate picture of the fractions of fossil and biogenic carbon. The estimated empirical values of materials of fossil origin are used to provide the estimated fossil carbon content of household, bulky waste, and commercial wastes (Table 2).

As shown in Table 2, textiles in household waste consist mostly of biogenic materials like cotton, and the fossil mass fraction is only 35%. Bulky waste contains a lot of carpets made of synthetic materials, and the fossil carbon fraction is 80%. Table 3 shows the calculated total carbon content of various materials.

There are no significant differences among the total carbon contents of the different waste materials. Thus, the same total carbon content is assumed for all waste types, whereas the total carbon contents vary significantly among the waste fractions as shown in Table 3. The calculated fossil carbon contents of various waste materials and three different types of wastes (household, bulky, and commercial) are shown in Table 4.

The fossil carbon content is 62%; therefore, plastics bring most of the fossil carbon into the MSW. The fossil carbon contents of all other waste fractions are considerably lower and they range from 3% to 36%.

The estimated content of fossil carbon in the waste fraction multiplied by the corresponding wet weighted amount in the waste composition (see Table 1) yields the content of fossil carbon in each waste material of the three types of wastes (Table 5).

Table 5 shows that, in total, 8.8% of the wet household waste consists of fossil carbon, 12.1% of the bulky waste, and 12.4% of the commercial waste that is similar to household waste.

During combustion, the carbon is oxidized to CO_2 and emitted to the atmosphere. A small amount of carbon may not be oxidized and leaves in the incinerator ash. In the literature, the oxidation of carbon, that is, the efficiency of combustion, is reported to be in the range of 95% and 100% [10, 12, 14]. In the following calculations, an oxidation of 97% was assumed for all waste types. Also, the conversion factor from carbon to CO_2 , that is, 44/12 on the basis of the respective molecular weights, must be taken into account. On the basis of these data, the CO_2 emission factors for the three types of waste included in MSW are calculated to be as follows:

- Household waste: 0.312 Mg CO_2 /Mg waste
- Bulky waste: 0.429 Mg CO_2 /Mg waste
- Commercial waste similar to households: 0.443 Mg CO_2 /Mg waste

As noted earlier, in addition to CO_2 , the other important GHG gas emitted by WTE plants is nitrous oxide (N_2O). N_2O has a long residence time in the atmosphere of 114 years because there appears to be no natural removal processes for this gas [15]. Other GHG trace gases, such as carbon monoxide (CO) that has a global warming potential of 1.9 for a time horizon

Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 1 Composition of different waste types [10]

Waste fraction	Household waste (% ww)	Bulky waste (% ww)	Commercial similar household waste (% ww) (M.-%)
Organic	30.9	0.6	13.2
Wood	1.9	42.6	6.3
Textiles	4.9	5.3	3.0
Minerals	4.6	1.7	4.8
Composite materials	4.7	26.3	8.6
Hazardous materials	0.6	0.1	0
Others	10.6	11.0	7.3
Fine fraction <10 mm	14.7	0.2	17.5
Fe/NE metals	2.7	5.0	3.0
Paper/cardboard	10.5	2.4	17.1
Glass	4.9	0.1	4.4
Plastic	9.2	4.7	14.8

Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 2 Estimated fossil carbon fraction of materials contained in three types of waste [10]

Waste fraction	Household waste (% ww)	Bulky waste (% ww)	Commercial similar household waste (% ww) (M.-%)
Textiles	35	80	70
Composite materials	70	60	40
Others	20	90	50
Fine fraction <10 mm	40	40	20
Plastic	100	100	100

Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 3 Total carbon content of waste fractions with fossil portions [10]

Waste fraction	Total carbon content (Mg C/Mg of waste material)
Textiles	0.37
Composite materials	0.38
Others	0.19
Fine fraction <10 mm	0.14
Plastic	0.62

Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 4 Fossil carbon content per ton of waste, for different waste materials and type of wastes

Waste fraction	Household waste, Mg C _{fossil} /Mg waste	Bulky waste, Mg C _{fossil} /Mg waste	Commercial similar to household waste, Mg C _{fossil} /Mg waste
Textiles	0.13	0.30	0.26
Composite materials	0.26	0.22	0.15
Others	0.36	0.17	0.09
Fine fraction <10 mm	0.06	0.06	0.03
Plastic	0.62	0.62	0.62

of 100 years, are not considered because their contribution to GHG is negligible.

The estimated total GHG emissions for the three types of wastes collected in the MSW stream are shown below:

- Household waste: 0.315 Mg CO_{2,eq}/Mg ww
- Bulky waste: 0.432 Mg CO_{2,eq}/Mg ww
- Commercial waste similar to households: 0.446 Mg CO_{2,eq}/Mg ww

As discussed earlier, the identified global warming factors (GWF) depend on the waste composition and their biogenic and fossil carbon contents. The GWF can

vary considerably from area to area. For household waste in industrial countries, the GWF ranges between 0.253 Mg CO_{2,eq}/Mg waste [16] to 0.557 Mg CO_{2,eq}/Mg waste [17].

Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 5 Fossil carbon content per ton of waste type, for different waste materials and type of wastes

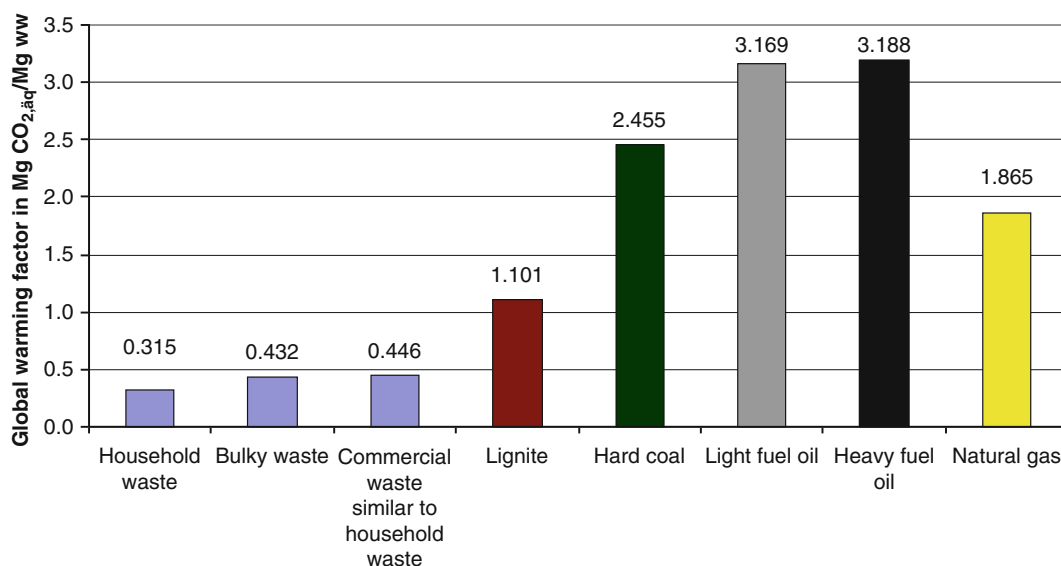
Waste fraction	Household waste, Mg C _{fossil} /Mg waste type	Bulky waste, Mg C _{fossil} /Mg waste type	Commercial similar household waste, Mg C _{fossil} /Mg waste type
Textiles	0.0063	0.0157	0.0077
Composite materials	0.0121	0.0570	0.0130
Others	0.0038	0.0188	0.0069
Fine fraction <10 mm	0.0082	0.0001	0.0049
Plastic	0.0571	0.0291	0.0917
Total	0.0876	0.1207	0.1244

Comparison of Global Warming Factors of MSW with Fossil Fuels

The carbon content of fossil fuels is by definition of fossil origin. Figure 4 compares the calculated global warming factors of three types of wastes in the MSW stream with those of fossil fuels listed in the Global Emission Model for Integrated Systems (GEMIS) [18].

Figure 4 shows that, with the exception of lignite, the global warming factors of MSW are one order of magnitude lower than the various types of fossil fuels used to generate electricity and heat. However, it is also necessary to consider the amount of energy that is generated by the various fuels. Table 6 shows the respective lower heating value (LHV) of the fuels.

The three waste types have a LHV ranging from 2.5 to 3.4 MWh/Mg of waste. The better fossil fuels range in LHV from 7 to 11.8 MWh/Mg of fuel, which is two to three times higher than MSW. The only exception is lignite which, because of its high water content, is in the same range as MSW. The LHV of natural gas is shown in MWh/Nm³ because at standard conditions (293 K and 101 kPa) it is in the gaseous form. The relation of the global warming factors and the corresponding LHV of the various fuels as used is shown in Fig. 5.



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 4

Calculated global warming factors of different waste types and of fossil fuels (per Mg of wet fuel)

Figure 5 shows that the GWF of MSW, per MWh of contained thermal energy, is only one third to one half that of fossil fuels. The main reason is that the biogenic carbon content of MSW has been assigned zero global warming factor, $\text{CO}_{2,\text{eq}}/\text{MWh}$.

The consequence is that when wastes are incinerated with the same energy efficiency like fossil fuels, the corresponding GHG emissions for the same amount of

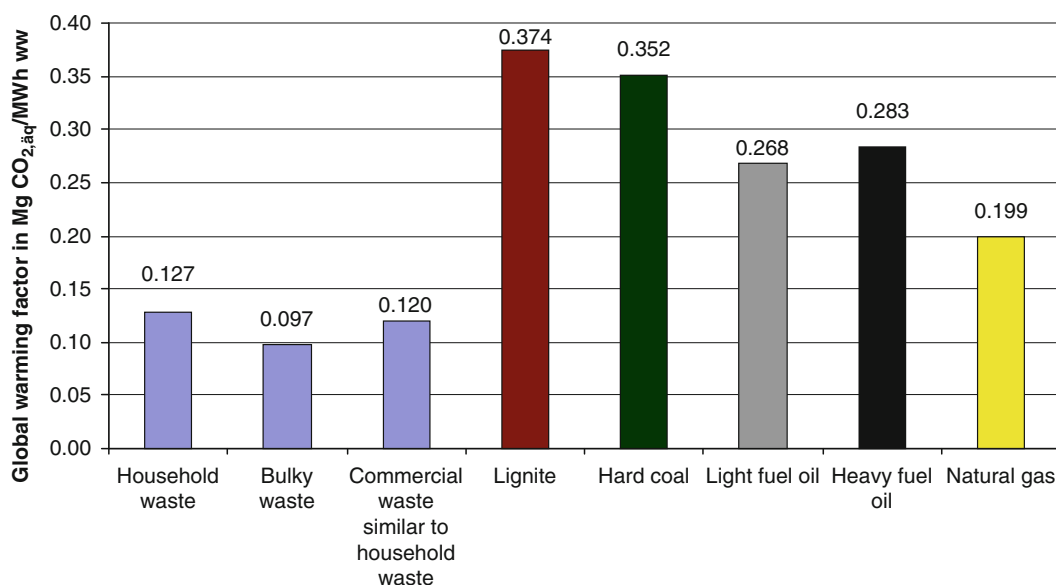
energy delivered is two to three times lower than for fossil fuels. Or, in other words, the released GHG emission for a delivered amount of energy, from waste or from natural gas, would be the same even if the efficiency of energy recovery of waste incineration were to be one half that of natural gas combustion.

The weighted average *gross* heat efficiency of German WTE plants in 2007 was estimated to be 36.5% [10] and the average weighted *net* heat efficiency was 27.8% [19]. For the 231 European plants that generate both electricity and district heating, an average weighted *gross* heat efficiency of 46% was calculated [20] but the net efficiency was not provided; on the basis of operating experience of WTE plants, it is expected that the average net heat efficiency of these plants is about 35%. The electrical efficiency does not differ significantly between countries because the technology for power generation is everywhere available, and even in developed countries like Germany old incinerators continue to operate.

In the case of heat generation efficiency, a difference between southwest and northern Europe is visible. The gross heat efficiency in southwestern Europe is around 22% and in northern Europe around 83% [21]. The principal reason is the extensive use of WTE plants for district heating in the north. Heat has to be used close

Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 6 Estimated LHV of selected fuels [10, 18] and own calculations

Fuel	LHV in MWh/Mg wet weight (MWh/Nm ³ for natural gas)
Household waste	2.468
Bulky waste	4.441
Commercial similar household waste	3.701
Lignite	2.941
Hard coal	6.980
Light fuel oil	11.836
Heavy fuel oil	11.275
Natural gas	9.392



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 5

Global warming factors of various types of wastes and fuels expressed as Mg CO₂/MWh of thermal energy

to the place of production because the transfer over higher distances is not economic. If there is no district heating network or industrial plant requiring heat, the low-level thermal energy exiting the turbine is wasted by air or water condensation. Because of their high emissions, decades ago, many incinerators were built far away from settlement structures. Consequently, even in Germany where energy efficiency is an important part in the GHG reduction targets of the government, the heat efficiency of the WTE plants is rather poor.

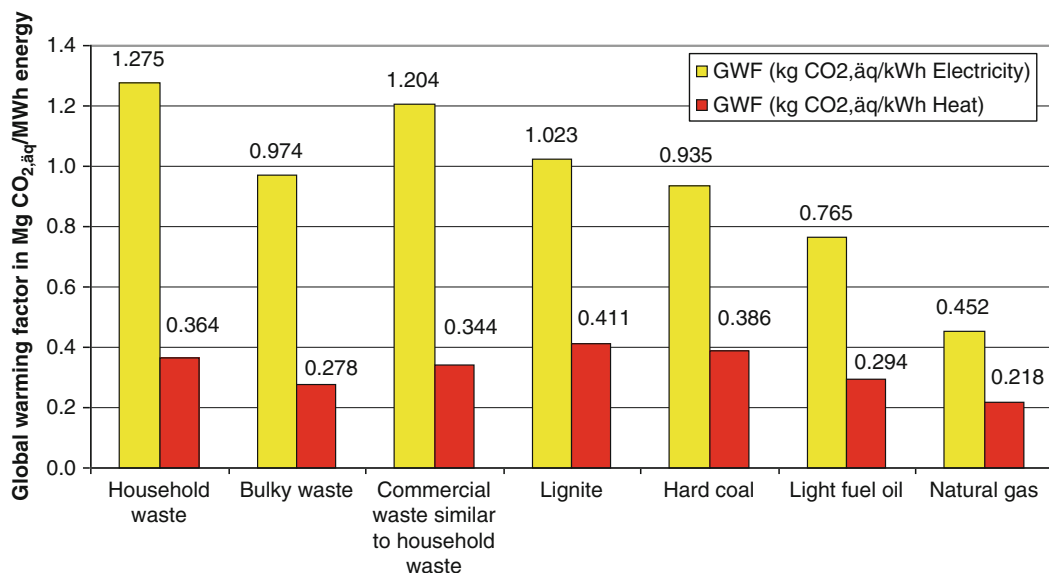
Another important fact is the climate. In south-western Europe heat is simply not needed because of relatively high temperatures all year round. A conversion from heat by absorption refrigeration to cooling is technically possible, but special cooling networks and cooling energy costumers are necessary. The opposite is in northern Europe. Cold climate and high heat requirements combined with an intelligent energy provision and generation system makes high efficiencies possible. In northern European countries, such as Denmark, WTE plants are positioned close to district heating networks and there is intelligent integration of the WTE plants in the heating system of an area. The simultaneous production of electrical power and heat in a single process of a power plant is called

combined heat and power generation (CHP) and ensures maximum efficiency in recovering energy. The very efficient incineration plants in northern Europe are using this technology.

Even when the combined heat and power generation is very efficient is financially supported, as for example in the case of CHP in Germany, most of the large fossil fuel power plants produce exclusively electricity. The electrical net efficiency of lignite power plants is around 36.6%, of hard coal power plants 37.6%, of light and heavy fuel oil around 35%, and of natural gas 43.9% [22]. Usually the produced heat of these power plants is not used, and in the case of lignite, hard coal and natural gas power plants that seek to have a high electrical efficiency, the steam is condensed to relatively low temperatures. This low-level energy cannot be used for district heating.

In comparison, smaller heating plants that are designed to produce only heat have a net heating efficiency of around 91%, using the same fossil fuels [23]. Because of high emissions, heavy fuel oil is not used anymore for power generation in most industrial countries. However, it is still used for fuelling ships and tankers.

Figure 6 shows the global warming factors per unit of energy content, as illustrated in Fig. 5, divided by the



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 6

Calculated global warming factors of different waste types and fossil fuels referring a unit of delivered energy

corresponding net efficiencies of the power plants using these fuels. The result of this division yields the amount of released GHG per unit of delivered electrical power or heat.

The low electrical net efficiency of 10% and the heat net efficiency of 35% from WtE plants cause the GWF of around 1 Mg CO_{2,eq}/MWh of electricity and about 0.3 Mg CO_{2,eq}/MWh of heat from wastes (Fig. 6). The higher net efficiencies of fossil-fired power plants of about 40% for electricity and 90% for heat result in GWF ranging from 0.45 to 1 Mg CO_{2,eq}/MWh of electricity and from 0.2 to 0.4 Mg CO_{2,eq}/MWh of heat for the most efficient power plants (Fig. 6).

However, it should be noted that the global warming factors for the most efficient WTE facilities are obtained by cogeneration of power and heat. The net efficiencies of 40% for power and 90% for heat for fossil-fired power plants are not possible in the case of cogeneration of electricity and heat.

Reduction of GHG Emissions and Conservation of Fossil Fuels by Means of Energy Recovery from Wastes

The energy released by the combustion in the waste incineration plants is used to produce electricity and/or heat. The effect of this is to decrease the consumption of fossil fuels, thus conserving a nonrenewable resource. The effect of this contribution to the overall mix of power generations differs from country to country. The amount of nuclear energy and renewable energy used, the amount of fossil fuels used, and the degree of cogeneration play an important role. Also, the efficiency of electricity generation from fossil fuels has a big influence on the electricity mix factor. In turn, the electricity mix factor results in the overall GWF (Mg CO_{2,eq}/MWh) of a nation. This factor is also affected by the upstream and downstream emissions associated with the mining, processing, and transport of the fuels; the construction and demolition of the energy-producing facilities; and the disposal of by-products, such as coal ash and nuclear wastes.

The electricity mix factors of many countries are calculated every year and are available in the literature. In the EU, the electrical mix factor ranges from 0.007 Mg CO_{2,eq}/MWh_{el} for Norway, which uses renewable hydropower, to 1.13 Mg CO_{2,eq}/MWh_{el} for Poland,

which depends on lignite-fired power plants [24]. Developed industrial economies with a wide mix of nuclear, coal, and renewable energy have GWF of about 0.6 Mg CO_{2,eq}/MWh_{el} [24, 25].

The heat mix factors are more difficult to calculate. The big difference between electricity and heat is in the ways that these two kinds of energies are produced and supplied to the users. Electricity is usually produced in a relatively small number of large power plants and is distributed by means of national grids. Heat is generated by a multitude of generating sources, including residential boilers, industrial and municipal boilers, and many other sources. Heat can be produced in cogenerating plants, as a by-product of industrial processes, and also by geothermal or solar energy and is delivered either as steam or warm water. It is therefore much more difficult to compile statistics on the GWF of heat generation than in the case of electricity. For the current 27 countries of the European Union, a weighted heat mix factor of 0.27–0.32 Mg CO_{2,eq}/MWh_{th} was estimated by Kreissig and Stoffregen [26] and Skovgaard et al. [27]. The German heat mix factor was calculated to be 0.216 Mg CO_{2,eq}/MWh_{th} [22] and for Greece to 0.468 Mg CO_{2,eq}/MWh_{th} [7].

As discussed earlier, the biogenic fraction of MSW energy is over 50%, and the energy delivered by waste incineration plants primarily substitutes energy from fossil fuels. Therefore, it is reasonable to calculate electrical and heat mix factors based on the energy provision of fossil fuels, that is, taking out of the energy mix factor of a country nuclear and renewable energy. The German federal environmental energy prefers to use the fossil energy mix factors for calculating the GHG emission savings resulting from MSW and other biomass combustion. In countries like Poland where nearly 100% of the energy is generated from fossil fuels, the fossil energy mix factors are nearly the same as the actual energy mix factors. In the case of Germany, 30% of electricity is generated by lignite, 60% by hard coal, and 10% by natural gas, when nuclear and renewable energy are excluded from the calculation of the energy mix; in this case, the fossil fuel mix factor is 0.886 Mg CO_{2,eq}/MWh_{el} versus the actual factor of 0.6 Mg CO_{2,eq}/MWh_{el} [22]. The German heat mix factor increases from 0.216 to 0.232 Mg CO_{2,eq}/MWh_{th} [22] when heat is exclusively generated by fossil fuels. When a nation obtains most of its energy from

natural gas, the mix factor is about 0.45 Mg CO_{2,äq}/MWh for electricity and 0.22 CO_{2,äq}/MWh for heat (Fig. 5). When the only fossil fuel is lignite coal, the fossil energy mix factor for electricity is about 1 Mg CO_{2,äq}/MWh and for heat 0.4 Mg CO_{2,äq}/MWh.

The third scenario is developed to examine the effect of energy provided by WTE as base load operation, that is, on a 24-h basis. Base load plants are primarily nuclear, lignite, and, in some cases, hard coal fired power plants; at times of high demand of electricity, the required peak load power is provided by power plants fired by light oil or natural gas. Waste incineration plants burn waste and provide energy continuously, 24 h a day, and, therefore, are a good base load provider.

In the case of Germany, base load electricity is primarily provided by lignite coal power plants with an electricity factor of 1,088 Mg CO_{2,äq}/MWh_{el} [22]. In Norway or France, where base load electricity is produced by renewable or nuclear energy, the corresponding factor is zero. Because of the difficulties involved in compiling heat generation data, as discussed earlier, it is not possible to calculate the base load heat factor; in this case, the fossil heat mix factor is assumed as the base load factor for generation of heat. The above discussion of GWF is summarized in Table 7.

As an example, in Germany the use of WTE reduces in the fossil energy mix scenario GHG emissions by the factor 0.886 Mg CO_{2,äq}/MWh, of electricity provided by the WTE, and 0.232 Mg CO_{2,äq}/MWh, for heat generated by the WTE. For the LHV of household waste presented in Table 6 and the power (10%) and

heat (35%) net efficiencies of waste incineration plants, 0.247 MWh of power and 0.864 MWh of heat are provided by the combustion of 1 t (1 Mg) of household wastes. Multiplying these amounts of energy by the emission factors of the fossil energy mix scenario yields the factor of 0.419 Mg CO_{2,äq}/Mg of household waste combusted. Comparison with the released GHG emissions from the combustion of household waste of 0.315 Mg CO_{2,äq}/Mg (Fig. 4) shows that there is a net reduction of 0.104 Mg CO_{2,äq}/Mg of MSW. Figure 7 shows the net prevention of GHG emissions for the three waste types and for the three substitution scenarios.

Figure 7 shows that for Germany, there is a net reduction of GHG emissions, in all three scenarios. The net reduction ranges from 0.02 to 0.14 Mg CO_{2,äq}/Mg for household waste to 0.17 to 0.39 Mg CO_{2,äq}/Mg for bulky waste. The net reduction is 0.315 for household waste and 0.446 Mg CO_{2,äq}/Mg for commercial waste similar to that of households. The difference is due to the higher LHV of the commercial waste.

Resource Conservation by Savings of Fossil Fuels

Each unit of energy produced by WTE conserves a unit of fossil fuel energy. Furthermore, every ton of incinerated waste conserves a certain amount of fossil fuels. Using the LHV of Table 6 and the power and heat net efficiencies mentioned earlier, one can calculate the mass of fossil fuels conserved per ton of MSW combusted (Fig. 8).

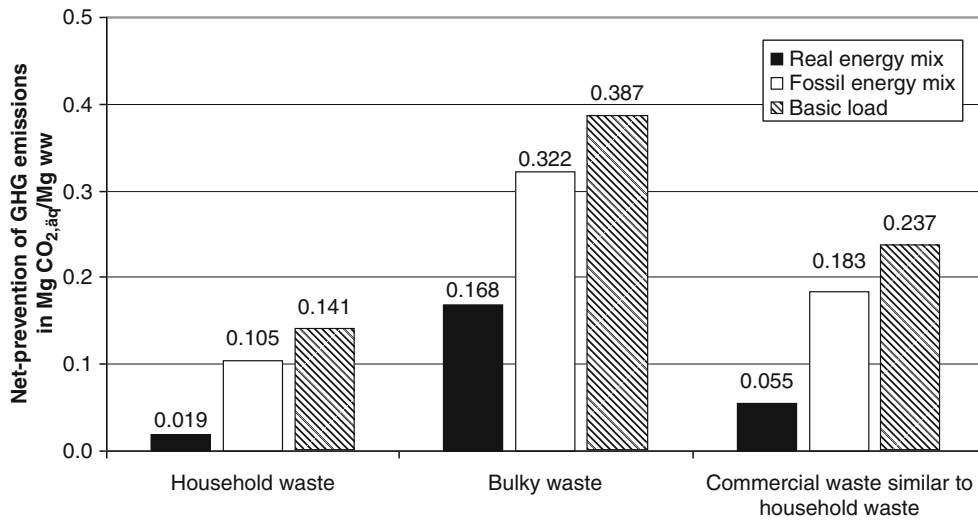
In an average WTE plant, between 0.15 Mg of natural gas and 0.55 Mg of lignite can be avoided by the incineration and energy recovery of 1 Mg of household waste (Fig. 8). In the case of heat generation, the amount of avoided fossil fuel per ton of MSW is higher because the energy efficiency of heat-generating WTE is higher than that for electricity. The higher is the LHV of the MSW combusted, the higher will be the reduction in GHG emissions. For example in the case of bulky waste (Table 6), 1 t of waste decreases the use of lignite by about 1 t.

Greenhouse Gas Emissions by Landfilling

As noted earlier, landfilling continues to be the dominant method of disposing MSW worldwide. Under the prevailing anaerobic conditions in a landfill and in the presence of water, the biogenic components of MSW

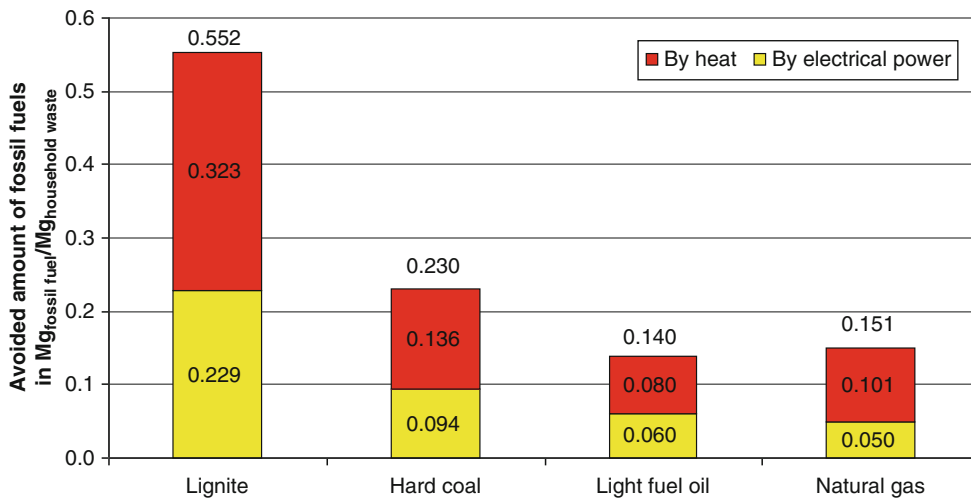
Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 7 Global warming factors for the generation of electricity and heat, under three substitution scenarios in EU

Substitution scenario	Electrical power [Mg CO _{2,äq} /MWh _{el}]	Heat [Mg CO _{2,äq} /MWh _{th}]
Actual energy mix	0.007–1.13	0.216–0.468
Fossil energy mix	0.45–1.13	0.22–0.5
Base load	0–1.13	0.216–0.468



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 7

Net reduction of GHG emissions in Germany for the three substitution scenarios and three types of waste



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 8

Conserved amount of various fossil fuels by the combustion of 1 t (1 Mg) of MSW in WTE

are converted to methane and carbon dioxide. Figure 9 illustrates that in Phase I, when wastes are first dumped, air can flow through the wastes; under these conditions, aerobic bacteria consume oxygen and break down the long molecular chains of complex proteins, lipids, and carbohydrates. The oxygen concentration decreases, and the concentration of produced carbon dioxide increases. Phase I lasts for days to several

months, depending on how loose or compressed the waste is. In Phase II, oxygen is depleted, and carbon dioxide reaches the concentration of over 60%. When anaerobic conditions are reached, methanogenic bacteria produce methane and carbon dioxide and, at the beginning of Phase IV, the generated landfill gas consists of nearly equal volumes of methane and carbon dioxide. It may take a few years until the stable Phase IV

is reached, depending on waste composition, moisture content, temperature within the landfill, and other factors. Phase IV can last several decades with a stable landfill gas composition.

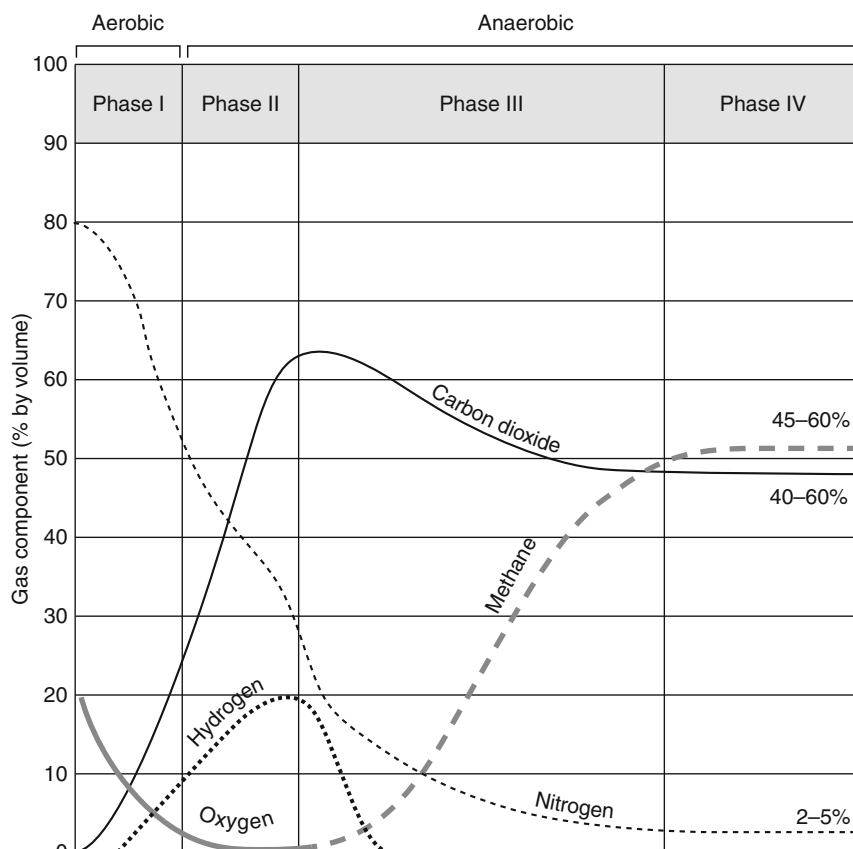
If there is no infiltration of rainwater in the landfill, the production of landfill gas decreases.

Modern, controlled landfills, also called sanitary landfills, are provided with ground sealing and also landfill gas collection systems. However, it may take several years for the landfills to be covered and the gas collection system to start operating. During this time, rainwater can flow through the wastes, and methane gas is formed. The amount of resulting GHG emissions from landfills is primarily dependent on the biogenic carbon and the water content in the wastes. The degradation of the biogenic carbon depends more on the water to solids ratio than on the age of the landfill.

In order to calculate the production of landfill gas, one needs to know the amount of biogenic carbon in the different wastes types. The amounts shown below are based on the calculation of the fossil carbon (see section on “[Global Warming Factors of different Waste Types in Germany](#)”):

- Household waste: 0.122 Mg C_{bio} /Mg waste
- Bulky waste: 0.232 Mg C_{bio} /Mg waste
- Commercial waste similar to household waste: 0.097 Mg C_{bio} /Mg waste

In practice, the amount of produced landfill gas is calculated by means of mathematical models that describe the decomposition and gas generation processes using mathematically formulated procedures in combination with empirical factors. A commonly used model in Europe was developed by Tabasaran and



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 9

Changes in landfill gas composition with time [28]

Rettenberger on the basis of the following formula [29]:

$$G_t = 1.868 \cdot C_{bio} \cdot (0.014 \cdot T + 0.28) \cdot (1 - 10^{-kt})$$

where

G_t = produced amount of landfill gas [$\text{m}^3/\text{Mg ww}$]

C_{bio} = biodegradable carbon [$\text{Mg C}/\text{Mg ww}$]

T = temperature [$^{\circ}\text{C}$]

k = degradation constant [a^{-1}]

t = time since deposition [a]

The factor 1.868 indicates the amount of landfill gas (in liters) stoichiometrically produced by 1 g of biogenic carbon. The temperature T in landfills is in the mesophilic range and is assumed to be 30°C . The degradation constant is calculated from the formula by $\ln 2/T_{1/2}$, where the average half-life ($T_{1/2}$) used by the authors is 18 years, and k is calculated to be 0.04. Easily biodegradable material like food wastes have a shorter half-lives than paper and wood that contain lignin; however, to simplify the calculation, the half-life of 18 years is assumed for all waste types. The amount of landfill gas produced annually, using the formula of Tabasaran and Rettenberger, is shown in Fig. 10.

According to Fig. 10, landfill gas production would be nearly finished after 50 years. The accumulated amount of landfill gas over the 50-year period is 158 Nm^3 per Mg of wet weight of household waste,

301 Nm^3 per Mg for bulky waste and 126 Nm^3 per Mg of commercial waste similar to households.

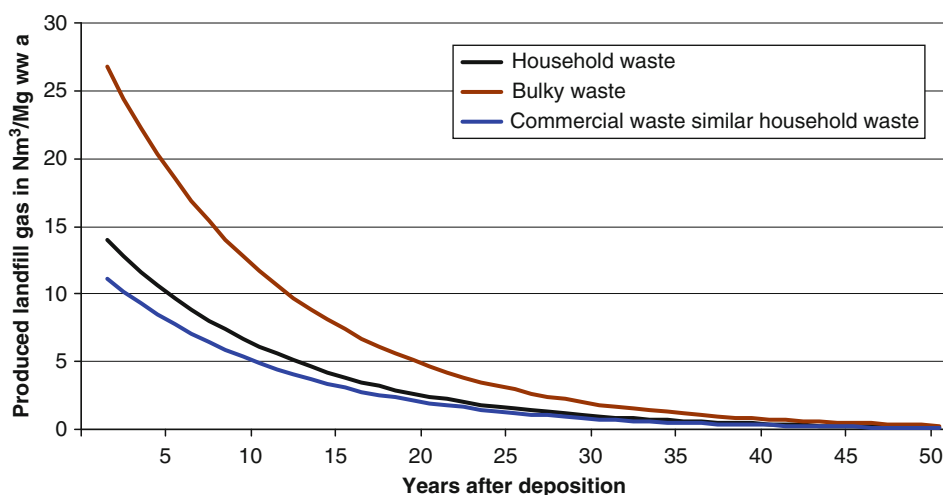
For an assumed distribution of 50% methane by volume and 50% carbon dioxide, the molar mass of methane (16 g/mol) and carbon dioxide (44 g/mol), the molar volume of gases at standard temperature and pressure (22.4 L/mol), and an assumed GWF for methane of 23, the following amounts of GHG emissions are released per ton of waste:

- Household waste: $1.449 \text{ Mg CO}_{2,\text{äq}}/\text{Mg waste}$
- Bulky waste: $2.766 \text{ Mg CO}_{2,\text{äq}}/\text{Mg}$
- Commercial waste similar to household waste: $1.155 \text{ Mg CO}_{2,\text{äq}}/\text{Mg}$

These amounts will be emitted to the atmosphere in landfills that are not provided with gas collection systems. In controlled (sanitary) landfills, over 50% of the total biogas generated can be captured and the rest is emitted to the atmosphere. The following section shows that the amount of GHG emissions from landfills is considerably higher than the GWF associated with the combustion of MSW in WTE.

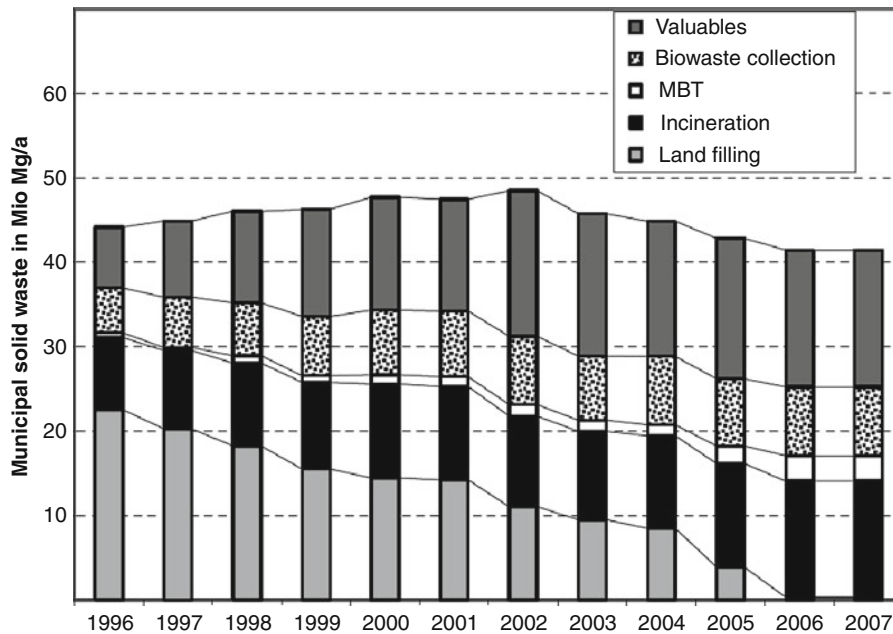
The Germany Example

Since 2005, landfilling of untreated MSW in Germany is prohibited; the primary reason is to reduce the landfill gas potential of the disposed waste.



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 10

Calculated amount of landfill gas generated annually



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 11

Disposition of MSW collected in Germany [31]

At this time, the post-recycling MSW goes directly to waste incineration or to mechanical and biological treatment (MBT) plants; in 2007, there was no landfilling of untreated MSW. In the MBT plants, some recyclables are recovered, green and food wastes are stabilized, and a high calorific value material is produced that may be used as a solid fuel (SRF or RDF). Figure 11 shows the disposition of municipal solid wastes in Germany, in the years 1996–2007 [30]. It can be seen that only a small fraction goes to BMT, 50% of the post-recycling MCW is combusted in WTE facilities, and in the period 1997–2007 landfilling was reduced from about 23 million tons to zero.

Greenhouse Gas Emissions by Waste Incineration in Germany 2007

In the year 2007, about 17.8 million tons of MSW were incinerated in 66 German waste incineration plants. The waste input composition was in general about 12.5 million Mg of household wastes, 0.5 million Mg of bulky wastes, and 4.8 million Mg of commercial wastes similar to household waste and other materials [10]. As discussed earlier, (see Fig. 4), the annual GHG emissions from household waste were 3.93, for bulky

waste 0.22, and for commercial waste similar to household waste 2.14 million Mg CO_{2,äq}.

For the calorific value of the waste mix of about 2.8 MWh/Mg [19], the thermal energy generated by the 17.8 million tons of MSW combusted in 2007 was 49.45 million MWh. In addition, an estimated [19] 1.62 million MWh of external energy was used in the form of light fuel oil used for start-up and shutdown operations or as natural gas used for reheating of the flue gas prior to selective catalytic reduction (SCR); Bilitewski [10] determined that 24% of this auxiliary energy was provided by light fuel oil and 76% by natural gas; as stated earlier (Fig. 4), the GWF of light fuel oil is 0.268 Mg CO_{2,äq}/MWh and of natural 0.199 Mg CO_{2,äq}/Mg. This use of auxiliary fuels resulted in the GHG emission of 0.35 million Mg CO_{2,äq}. In total, by waste incineration in Germany in the year 2007 emitted about 6.64 million Mg CO_{2,äq}.

Reduction of Greenhouse Gas Emissions in Germany by Means of Waste Incineration in 2007

Table 8 shows the amounts of electricity and thermal energy provided by the 66 German waste incineration plants in 2007.

Greenhouse Gas Emission Reduction by Waste-to-Energy. Table 8 Energy generated by all WTE plants of Germany in 2007 [19]

Type of energy	Amount of energy provided, in million MWh/a
Electrical power	5.16
Heat	13.75
Total	18.91

Since the thermal energy input was estimated earlier to be 49.45 million MWh/a minus the 1.62 million MWh of external fuel, the energy efficiency of delivered electrical power was 10.1% and the energy efficiency of delivered heat 26.9%. The combined efficiency of delivered electrical and thermal energy generation was 37%.

In determining the effect of WTE in substituting fossil energy, three scenarios were considered, same as discussed earlier:

- Actual German energy scenario
 - Power mix: nuclear 23%, natural gas 13%, lignite 23%, hard coal 20%, renewables 15%, others 6%; electricity GWF of 0.596 Mg CO_{2,äq}/MWh_{el} [22]
 - Heat mix (39% hard coal, 42% natural gas, 12% lignite, 6% waste); heat GWF of 0.216 Mg CO_{2,äq}/MWh_{th} [32]
- Fossil fuel scenario
 - Power mix: 60% hard coal, 30% lignite, 10% natural gas; electricity GWF of 0.886 Mg CO_{2,äq}/MWh_{el} [22]
 - Heat mix: 57% natural gas, 40.5% light fuel oil, 2.5% coal; heat GWF of 0.232 Mg CO_{2,äq}/MWh_{el} [22, 33]
- Base load scenario
 - Electricity factor based on lignite coal; GWF of 1.088 Mg CO_{2,äq}/MWh_{el} [22]
 - Heat factor based on the actual German heat mix; heat GWF of 0.216 Mg CO_{2,äq}/MWh_{th} [32]

The actual German power mix consists of a wide range of energy sources, so as to not to be dependent on one source. The future aim is to reduce the nuclear power fraction and replace it by renewable energy. As was shown in Fig. 5, natural gas is the “cleanest” fossil energy with the highest conversion efficiencies.

Clean is not only meant in terms of low emission of carbon dioxide; during combustion of natural gas, other emissions such as sulfur oxides, hydrochloric acid, carbon monoxide, and particulates are much lower than from other fossil fuels. For this reason, Germany is making further efforts to increase the use of natural gas by expanding gas pipeline connections to Russia.

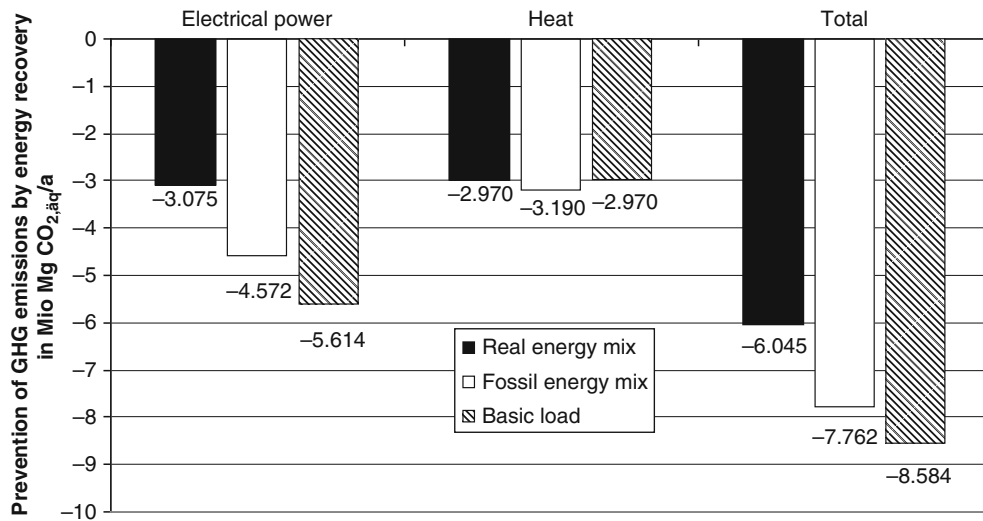
As noted earlier, data on industrial heat is hard to obtain; therefore, the above distribution of heat for the actual German energy scenario is based on the provision of district heat. The majority of provided heat is produced by hard coal and natural gas fired power plants, in combined heat and power generation. Actually, nearly 90% of the district heating in Germany is produced in cogeneration with electricity [32]. This cogeneration is financially supported by the German government by implementing high-efficiency technologies and in order to reduce GHG emissions.

The fossil power mix for Germany consists of hard coal and to a lesser degree by lignite and natural gas. The fossil heat mix is based on the energy source share of the German heat market in 2003 [32].

Apart from nuclear power, the base load of electricity in Germany is provided by lignite. Nuclear power has little effect on GHG, so the base load scenario involves substitution of lignite by MSW incineration. For the base load of heat, the actual heat mix factor was used, because the heat requirements change with season rather than time of the day.

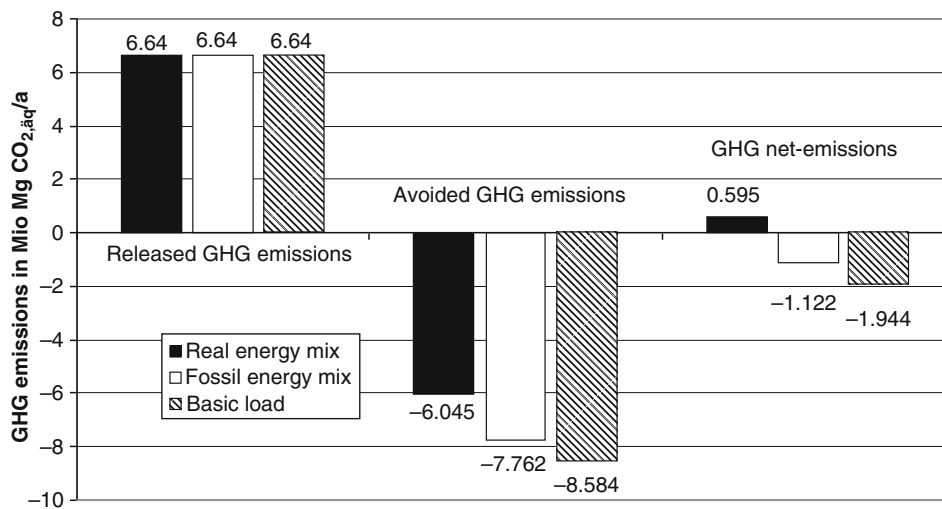
Multiplying the amounts of electricity and heat provided by WTE (see Table 8) by the substitution factors of the three scenarios yields the amounts of GHG emissions avoided by substituting MSW incineration for the combustion of fossil fuels (Fig. 12).

The avoided GHG emissions are shown as negative numbers in Fig. 12. The three bars on the left side of Fig. 12 show the effect of WTE in substituting fossil fuels for electricity generation. Substitution of heat energy (three bars in the middle) results in lower avoidance of GHG emissions because heat generation in Germany is very efficient, and the substitution factors lower. The avoided GHG emissions range from 3 and 5.6 million Mg CO_{2,äq}/a. In total, the GHG emissions avoided by WTE electrical and thermal energy range from 6 to 8.6 Mio Mg CO_{2,äq} per year. By adding the WTE GHG emissions and the avoided (i.e., negative)



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 12

Avoided GHG Emissions in Germany by means of energy recovery in waste incineration plants in 2007, under three energy mix scenarios



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 13

GHG emissions, avoidance of emissions, and net emissions of German waste incineration plants in 2007

emissions of fossil fuels avoided because of the energy generated by WYE (Fig. 12) results in the net GHG emissions shown in Fig. 13 (last three bars).

The German example shows that the scenario chosen for substitution of energy is very important and has a real impact on the GHG balance of WTE plants. In

the case of substitution in the actual German energy mix, the GHG net emissions (0.6 million Mg CO_{2,äq}/year) is positive. For the substitution of the fossil energy mix and the base load scenarios, the GHG balance is negative, and GHG emissions of 1.1 and 2 million Mg CO_{2,äq}/year are avoided.

Comparison to Landfilling

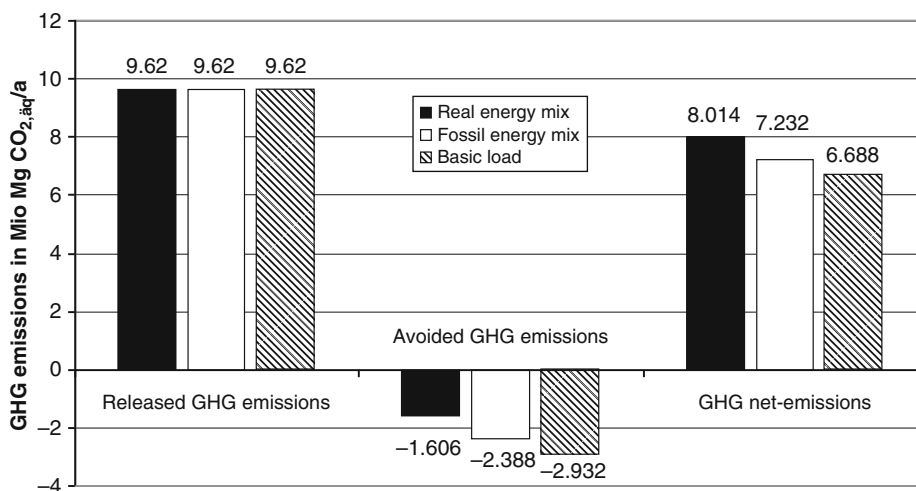
Currently, 95% of all German landfills are equipped with a gas collection system, and the biogas collection efficiency has been estimated at 60% [33]. If it is assumed that the 17.8 million Mg of incinerated MSW were landfilled and using the calculations presented in the section “[Greenhouse Gas Emissions by Landfilling](#)”, 9.6 million Mg CO_{2,äq} would be directly released. The gas collection systems would capture about 0.4 million Mg of methane. Since the LHV of methane is 13.9 MWh/Mg, the captured gas would provide about 5.6 million MWh that could be used to generate energy in the form of electricity and heat. Usually the landfill gas is burned in slightly modified gasoline engines to produce electricity only. Heat is generally not needed at landfills, and therefore not recovered. The electrical efficiency of the gasoline engines is around 35%. Assumed that all of the captured landfill gas can be used in gasoline engines, about 2 million MWh of electricity would be produced. This electricity would substitute energy produced by other fossil fuels, and thus avoid GHG emissions. On the basis of the substitution scenarios for Germany (see Section Prevention of Greenhouse Gas Emissions by Energy Recovery in German Waste Incineration Plants) between 1.6 and 3 million Mg CO_{2,äq} of fossil fuel emissions would be avoided. Figure 14 shows

what would be the effect of landfilling, instead of combusting, 17.8 million Mg on the GHG emissions balance for Germany, for the same three energy mix scenarios.

Figure 14 shows that even if all of the captured methane were used to generate electricity, the GHG net emissions would be between 6.7 and 8 million Mg CO_{2,äq}. In comparison to waste incineration (see Fig. 13), the GHG net emissions by land filling would be about 8 million Mg CO_{2,äq}, higher than in the case of waste incineration with energy recovery. This amount corresponds to about 0.5 t of carbon dioxide per ton of MSW landfilled rather than combusted. The landfill emissions would be much greater for landfills that do not recover methane or flare the landfill gas instead of using it to generate electricity.

Future Directions

This entry shows clearly that waste incineration results in much lower GHG emissions than landfilling. One of the reasons is that only CO₂ is emitted during incineration, while the landfill gas that is not captured consists of about 50% methane that has a global warming potential that is 23 times higher than the same volume of carbon dioxide. The second reason is that much more energy is generated by the combustion of MSW



Greenhouse Gas Emission Reduction by Waste-to-Energy. Figure 14

GHG emissions in Germany if 17.8 million tons of MSW were landfilled rather than combusted in 2007

than from the combustion of the methane generated and captured during landfilling.

Many communities have begun to collect biodegradable materials separately. The result is less biogenic fixed carbon in the household wastes and a higher amount of plastics and composite materials with high amounts of fossil-fixed carbon. The result are a lower amount of landfill gas generated during landfilling and a corresponding reduction in GHG emissions. For the waste incineration, removal of organics results in a higher concentration of fossil-based wastes and higher GHG emissions. It also means higher calorific value of the MSW which should result in higher thermal efficiency. In any case, it is necessary for WTE plants to strive to increase their thermal efficiency and energy recovery. This will result in higher amounts of produced and delivered energy and, thus, higher substitution of fossil fuels and increased avoidance of GHG emissions.

Bibliography

Primary Literature

1. National Inventory Report for the German Greenhouse Gas Inventory 1990–2007, Submission under the United Nations Framework Convention on Climate Change and the Kyoto Protocol 2008, Federal Environment Office, Dessau-Roßlau, April 2008, p 47
2. Metz B et al (eds) (2005) Special report on safeguarding the ozone layer and the global climate system: issues related to hydrofluorocarbons and perfluorocarbons. Cambridge University Press, Cambridge/New York
3. Emission control: Energy conversion in thermal solid waste treatment. VDI 3460 Part 2, 2007, issue German/English, ICS 13.030.40, 27.190, p 13
4. Guendehou G, Koch M, Hockstad L, Pipatti R, Yamada M (1997) Incineration and open burning of waste. In: IPCC guidelines for national greenhouse gas inventories, vol 5, chap 5, p 5.5
5. Obermoser M, Fellner J, Rechenberger H (2009) Determination of reliable CO₂ emission factors for Waste-to-energy plants. Waste Manage Res 27(9):407–413, Applied greenhouse gas accounting: methodologies and cases
6. Riber C, Pedersen C, Christensen TH (2009) Chemical composition of material fractions in Danish household waste. Waste Manage 29:1251–1257
7. Gentil E, Clavreul J, Christensen TH (2009) Global warming factor of MSW management in Europe. Waste Manage Res 27(9):850–860, Applied greenhouse gas accounting: methodologies and cases
8. Manfredi S, Scharff HM, Tonini D, Christensen TH (2009) Landfilling of waste: accounting of GHGs and GW contributions. Waste Manage Res 27(8):825–836, Fundamental in greenhouse gas accounting: concepts and mechanisms
9. Astrup T, Moeller J, Fruergaard T (2009) Incineration and co-combustion of waste: accounting GHG and global warming contribution. Waste Manage Res 27(8):789–799, Applied greenhouse gas accounting: methodologies and cases
10. Bilitewski B, Wunsch C, Jager J, Hoffmann M (2010) Energieeffizienzsteigerung und CO₂-Vermeidungspotenziale bei der Müllverbrennung – technische und wirtschaftliche Bewertung. EdDE-Dokumentation 13, Entsorgungsgemeinschaft der deutschen Entsorgungswirtschaft e.V., April 2010, p 16
11. Roland C, Scheibengraf M (2003) Biologisch abbaubarer Kohlenstoff im Restmüll. Umweltbundesamt, Berichte BE-236. Wien, 2003
12. Pipatti R, Sharma C, Yamada M, Alves J, Gao Q, Guendehou G, Koch M, López Cabrera C, Mareckova K, Oonk H, Scheehle E, Smith A, Svardal P, Vieira S (2006) Waste generation, composition and management data. Solid waste disposal. In: IPCC guidelines for national greenhouse gas inventories, vol 5, chaps 2 and 3
13. Dehoust G, Schüler D, Vogt R, Giegrich J (2010) Klimaschutzpotenziale der Abfallwirtschaft. IFEU und Ökoinstitut e.V, Darmstadt/Heidelberg/Berlin
14. Hiraishi T, Nyenzi B, Miguez J, Alves J, Boeckx P, Brown K, Hoppaus R, Jubb C, Kerr T, Kleffhelgaard T, Lucon O, Mauschitz G, Midaglia C, Milton M, Mondshine M, Oonk H, Paradiz B, Steczko K, Teixeira G, Towprayoon S, Yesserkepova I (2001) Waste. In: IPCC good practice guidance and uncertainty management in national greenhouse gas inventories, chap 5, p 5.29
15. Scheutz P, Kjeldsen P, Gentil E (2009) Greenhouse gases, radiative forcing, global warming potential and waste management – an introduction. Waste Manage Res 27(8):716–723, Fundamental in greenhouse gas accounting: concepts and mechanisms
16. Bilitewski B, Schirmer M, Niestroj J, Wagner J (2005) Ökologische Effekte der Müllverbrennung durch Energienutzung, EdDE-Dokumentation 10, Entsorgungsgemeinschaft der deutschen Entsorgungswirtschaft e.V. Pirna, 2005, p 24
17. Houghton JT et al (eds) (1996) Intergovernmental Panel on Climate Change: climate change 1995: the science of climate change. Contribution of working group I to the second assessment report of the intergovernmental panel on climate change, Cambridge University Press, Cambridge
18. Global Emission Model for Integrated Systems (2009) Version 4.6, Öko-Institut e.V
19. Treder M (2008) Energieerzeugung und Klimarelevanz der W-t-E Anlagen in Deutschland (Kurzfassung vom 16.07.2008), Würzburg
20. Reimann DO (2009) CEWEP Energy Report II (status 2004–2007). CEWEP, Bamberg

21. Fruergaard T, Ekval T, Astrup T (2009) Energy use and recovery in waste management and implications for accounting of greenhouse gases and global warming contributions. *Waste Manage Res* 27(8):724–737, Fundamental in greenhouse gas accounting: concepts and mechanisms
22. Staiß F, Linkohr C, Zimmer U, Musiol F, Ottmüller M (2008) Erneuerbare energien in zahlen, nationale und internationale entwicklungen. Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (BMU), Berlin, Juni 2008
23. BREF/BAT Waste Incineration for Integrated Pollution Prevention and Control (IPPC) (2005) "Draft Reference Document on the Best Available techniques for Waste Incineration" European Commission, EIPPC Bureau Sevilla, Final Draft, May 2005
24. Dones R, Heck T, Hirschberg S (2004) Greenhouse Gas emissions from energy systems, comparison and overview. In: Cleveland C (ed) *Encyclopedia of energy*, vol 3. Academic/Elsevier, San Diego, pp 77–95
25. Umweltbundesamt – German Federal Office for Environment, press information 34/2008, Dessau-Rosslau, 16.05.2008, Germany
26. Kressig J, Stoffregen A (2008) Life cycle assessment of waste-to-energy plants in Europe – modeling of thermal treatment of municipal and similar waste to calculate eco-profiles for the European reference life cycle data system (ELCD), Performed for CEWEP by PE International, Leihnfelden-Echterdingen, Germany, 2008
27. Skovgaard M, Heddal N, Valanueva A, Andersen FM, Larsen H (2008) Municipal solid waste management and greenhouse gases, ETC/RWM working paper 2008/1, European Topic Center (ETC) on Resource and Waste Management (RWM), Copenhagen, 2008
28. ATSDR (2001) Landfill gas basic, agency for toxic substances & disease registry. In: *Landfill gas primer – an overview for environmental health professionals*. Atlanta, chap 2, November 2001
29. Tabasaran O, Rettenberger G (1987) Grundlagen zur Planung von Entgasungsanlagen, Handbuch Müll und Abfall, Kennz. 4547, Lieferung 1/87, E. Schmidt Verlag
30. Abfallbilanz, Umwelt, Statistisches Bundesamt, Wiesbaden, erschienen im Juli 2010
31. National Inventory Report for the German Greenhouse Gas Inventory 1990–2008, Submission under the United Nations Framework Convention on Climate Change and the Kyoto Protocol 2010, Federal Environment Office, Dessau-Roßlau, April 2010
32. Fritsche U, Rausch L (2008) Bestimmung spezifischer Treibhausgas-Emissionsfaktoren für Fernwärme, Bereich Energie & Klimaschutz. Öko-Institut, Büro Darmstadt, Im Auftrag des Umweltbundesamtes, Dessau-Roßlau, Mai
33. Bundesministerium für Wirtschaft und Technologie (Federal Ministry for Economics and Technology) (2006) Energieversorgung für Deutschland, Statusbericht für den Energiegipfel am 3 April 2006. Berlin, March 2006, p 61

Groundwater Impacts of Radioactive Wastes and Associated Environmental Modeling Assessment

RUI MA^{1,3}, CHUNMIAO ZHENG¹, CHONGXUAN LIU²

¹Department of Geological Sciences, University of Alabama, Tuscaloosa, AL, USA

²Pacific Northwest National Laboratory, Richland, WA, USA

³China University of Geosciences, Wuhan, China

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Overview on Groundwater Impacts of Radioactive Wastes
 Controlling Biogeochemical Processes
 Environmental Modeling Assessment
 Future Directions
 Bibliography

Glossary

Absorption The process in which a dissolved substance is incorporated into the interior of a solid grain.

Adsorption The adhesion of a chemical species onto the solid surface.

Contaminant fate and transport The ultimate state of contaminants and the processes by which the contaminants migrate through the subsurface.

Groundwater Water that exists in liquid form beneath the land surface, filling the cracks, voids, and pore spaces in earth materials. The subsurface strata that store and transmit groundwater are referred to as aquifers.

Mass transfer The interaction and exchange of solutes in mobile state with those in immobile state through either physical or chemical processes.

Radioactive Waste A waste product that contains radioactive material. The majority of radioactive waste is "low-level" waste, which has low levels of radioactivity per unit of mass or volume.

Depending on the type and nature of radioactive wastes, it could take hours to thousands of years to diminish their radioactivity.

Reactive transport model The mathematical model that couples hydrogeological, geochemical, and biological processes to simulate and predict the contaminant fate and transport in the subsurface.

Sorption/desorption Sorption is a general term used to describe both adsorption and absorption by which a dissolved substance is attached to the surface or incorporated into the interior of a solid grain. The reverse process from the sorbed phase to the dissolved phase is referred to as desorption.

Definition of the Subject

Ever since the dawn of the nuclear age, especially since the 1970s with the work related to exploration of Yucca Mountain as a potential nuclear waste repository [1] and remediation of former nuclear fuel processing sites such as Hanford [2], there have been significant public concerns over groundwater impacts of radioactive wastes. This is because groundwater is a vital water resource with tremendous values to public water supplies and ecological lives, and because groundwater provides a pathway for potential spread of radioactive contaminants, posing significant risks to human health and ecological systems. Investigation of groundwater impacts of radioactive wastes requires understanding hydrogeological, geochemical, and biological processes that control the fate and transport of radionuclide contaminants. It will also require developing numerical simulators to integrate these processes for future projection under both natural and engineered remediation scenarios. While some sporadic information is available at specific sites, few attempts have been made to integrate scattered information into a coherent framework to answer relevant questions such as what are the most common sources of radioactive contaminants in groundwater and how they are transported and transformed in the aquifer.

Introduction

Radioactive waste contamination in soil and groundwater poses long-term risks to human health and environment. Public awareness on this subject has steadily

increased worldwide since the inception of the nuclear age. Significant efforts have been made to understand the fate and transport of radionuclide wastes in environments, and a number of research and remediation programs have been established to remediate contaminated sites in the USA.

The radionuclide fate and transport in groundwater is controlled by hydrologic, microbiologic, and geochemical processes that operate in the subsurface environment. These processes and their coupling control contaminant migration and persistence, and efficiency of remediation technologies. The understanding of the fate and transport of radionuclide contaminants can improve the ability to forecast contaminant destination and select cost-effective remediation technologies such as mobilization, immobilization, or in-ground degradation.

Reactive transport models are important tools to systematically integrate physical, chemical, and biological processes and data that are critical to understand and predict the fate and transport of radionuclide contaminants in the subsurface. Properly designed and calibrated models can describe the interactions of competing processes at a range of spatial and time scales, and hence are helpful for optimizing field operations and designing monitoring systems for remediating contaminated sites.

This entry is intended to provide a review of the major sources of radioactive wastes and their impacts on groundwater contamination, to discuss the major biogeochemical processes that control the transport and fate of radionuclide contaminants in groundwater, and briefly describe the evolution of mathematical models designed to simulate and assess the transport and transformation of radionuclides in groundwater.

Overview on Groundwater Impacts of Radioactive Wastes

Source of Radioactive Wastes

Radioactive wastes were generated primarily from the production of nuclear fuels for the weapons program and electrical energy, development and operation of commercial power reactors, nuclear weapons tests, fuel reprocessing, waste storage and disposal activities, and nuclear accidents. According to the statistics by Ahearne [3], the combined volume of all radioactive

wastes (excluding that in the soil and water) from both the government and commercial sources in the USA is about 5.5 million cubic meters, and the total radioactivity from all anthropogenic sources is about 31 billion Ci. Over the past decades, the uranium mining for production of nuclear fuels has resulted in a large volume of mine and mill tailings, which contain all of the naturally occurring radioactive elements. This has left a legacy of environmental pollution across the countries in the world, such as former Soviet Union, USA, Germany, France, and Eastern European countries [4].

Contamination of Groundwater Caused by Radioactive Waste Release

Radionuclide contaminants have been detected in subsurface sediments and groundwater as a result of intentional and accidental release of radionuclide-containing wastes during storage, processing, and disposal of nuclear materials [5–7] as well as the leaching of uranium mill tailings [8, 9]. Thirty to eighty million cubic meters of soil and 1,800–4,700 million cubic meters of water have been contaminated by radionuclides in USA [10]. Among them, most contamination occurred at US Department of Energy (DOE) sites used for nuclear weapons production, where totally over 5,700 individual contaminant plumes have been detected in subsurface [11]. At DOE sites, the radionuclides (e.g., uranium, cesium, strontium, thorium, and tritium) have been normally co-disposed with chlorinated solvents (e.g., perchloroethylene (PCE) and trichloroethylene (TCE)) and metals (e.g., lead, chromium, and mercury) and thus have been often identified together with these contaminants in groundwater [5]. According to DOE estimation, approximately 38 million cubic meters of groundwater was contaminated from the wastes generated from uranium mill processing or mill tailings [12]. Among the radionuclide contaminants, actinide elements uranium, neptunium, and plutonium are the most problematic [13] because of their long decay half-lives in subsurface sediments and groundwater.

Groundwater Contamination by Radionuclides at Well-Known Sites

Thousands of sites around the world are currently contaminated with radionuclides [7]. In particular,

spectacular examples of such sites can be found in the former Soviet Union and USA. Major contaminated areas are often located at or near facilities that reprocessed nuclear fuels from production reactors [6]. In the USA, the major sites are at Hanford in Washington state, Savannah River in South Carolina, and Oak Ridge in Tennessee [5, 6]. The contamination at Oak Ridge was caused by underground injection of cesium and strontium wastes, and at Savannah River resulted from the release of mixed fission product solutions into streams and seepage basins, and at Hanford originated from the discharge of mixed fission wastes into soils and surface ponds [6].

Among the contaminated sites at USA, the Hanford site has been dubbed “the dirtiest place on Earth” [14]. The contamination at the Hanford site, which was divided into three areas, namely, 100 Area, 200 East and West Areas, and 300 Area, has mainly occurred at locations of nuclear fuel fabrication, fuel irradiation, strategic radionuclide separation, and waste storage and disposal for plutonium production [5]. Sixty-seven of the 149 single shell tanks used for storing nuclear wastes were suspected to have released over 1.9 million liters of tank waste to the vadose zone [15]. The tank leaks, combined with discharge of liquid waste through ponds, trenches, pipelines, and cribs, caused a large quantity of radioactivity (through the year 2000: two million Ci) and 100,000–300,000 t of toxic chemicals residing in the vadose zone [2]. Wastes at the site have migrated through the vadose zone, resulting in groundwater contamination including nitrate, chromium, tritium, uranium, strontium-90, technetium-99, iodine, carbon tetrachloride, and others [16–18]. For example, waste disposal in 300 Area resulted in a groundwater plume of uranium (VI) with an area of 0.4–0.5 km² that exceeded the drinking water standard of 30 µg/L up to present [19].

For over 30 years, trillions of liters of acidic plating wastes containing high levels of uranium and nitric acid were generated at the Y-12 Facility, Oak Ridge, and Tennessee and were discarded into unlined S-3 Ponds. The wastes were neutralized and denitrified and the area was capped and converted to a parking lot in 1984 [20]. Despite these treatments, radionuclide contaminants continued to migrate from the source along geologic strike and dip to groundwater [21, 22].

The Savannah River Site was established in the early 1950s to produce nuclear materials, primarily tritium and plutonium-239 for nuclear weapon purposes, but also plutonium-238 and various transplutonium radionuclides for medical, industrial, and scientific applications. The production was ended in 1988 [23, 24]. The operations at the Savannah River Site have resulted in the migration of radionuclides (e.g., uranium, cesium, radium, thorium, and tritium) into groundwater at various locations, predominantly in the central areas of the site [24, 25].

Many abandoned mine processing sites were also contaminated with radionuclides during uranium mining activities. Contaminated groundwater at the former uranium mill site located at Naturita, Colorado, which processed uranium and vanadium ores intermittently from 1930s to 1958, occurred in the thin alluvial deposits of the San Miguel River floodplain. High concentrations of uranium were measured in the groundwater below and downgradient of the former tailings pile [26]. The uranium at the Old Rifle UMTRA field experiment site in western Colorado in the aquifer originated at mill tailings (now removed), percolated through a 4 m thick vadose zone to the water table, and was transported laterally through the aquifer via groundwater flow with maximum uranium concentrations of 300 $\mu\text{g/L}$ [27].

Controlling Biogeochemical Processes

The fate and transport of radionuclides in groundwater is controlled by geochemical and biological processes, in addition to the hydrological ones. This section discusses four biogeochemical processes that have been identified to have major effects on the fate and transport of radionuclide contaminants in groundwater.

Adsorption/Desorption

The adsorption to sediment surfaces is a major geochemical process controlling the mobility of radionuclide contaminants in oxic groundwater. This process is regulated by interfacial chemistry of the prevailing mineral surfaces [9, 26, 28–32]. For example, spectroscopic characterization, laboratory transport experiments, and numerical simulations have revealed that adsorption is a primary process controlling uranium transport in the groundwater at Hanford 300 Area [33, 34], at a U(VI) mill located in Naturita, Colorado

[26], at the Oak Ridge site in Tennessee [35], and at the Old Rifle UMTRA site in western Colorado [27]. Many experimental studies have also demonstrated that plutonium can be adsorbed onto a variety of minerals and mineral assemblages [36–38].

Sorption of actinides, particularly plutonium, onto submicrometer-sized colloids increases their mobility in groundwater [39]. Certain actinides can be stabilized in natural waters through the formation of actinide pseudo-colloids, in which the actinide sorbs onto aquatic colloids. This process alone can increase the actinide concentrations by many orders of magnitude over the values expected from solubility calculations [24, 39–41]. Colloid-facilitated transport is likely one of major mechanisms for long-distance transport of actinides in groundwater [39].

The kinetic adsorption/desorption behavior of radionuclide contaminants has often been observed in uranium (VI)-contaminated sediments with controlling mechanisms generally not well understood [32, 34, 42–47]. The kinetic uranium (VI) release from the Hanford 300 Area sediments was found to result from diffusional mass transfer from intragrain, intra-coating, and intragrain aggregate regions based on microscopic and spectroscopic characterizations of the sediments [48, 49].

Aqueous Complexation

Uranium (VI) can complex with various ligands such as carbonate in groundwater to form aqueous complexes. The aqueous complexation process stabilizes uranium in aqueous phase and decreases its tendency to bind to mineral surfaces [50]. Fox et al. [51] and Dong et al. [52] found that calcium can have a significant impact on the aqueous speciation of uranium (VI) under neutral to mild alkaline pH conditions through formation of ternary uranium (VI)-calcium-carbonate aqueous species. The aqueous and surface uranium (VI) complexation is sensitive to important groundwater chemical composition including pH, carbonate, and calcium concentrations [9, 26, 51]. Consequently, the hydrogeochemical conditions in groundwater at field sites have a great impact on radionuclide contaminant fate and transport. The groundwater redox conditions have also been found to impact the mobility of other selected radionuclides [53].

Precipitation/Dissolution

Radionuclide fate and transport is also affected by processes, such as precipitation/dissolution reactions [46, 54, 55], co-precipitation with other minerals [56], and microbially induced mineralization [22]. The actinide precipitation/dissolution may occur in intragrain regions where thermodynamic conditions for precipitation/dissolution reactions may significantly differ from bulk solution [46, 54].

Precipitation and co-precipitation processes play an important role in uranium stabilization under both reducing and oxidizing conditions [33, 57–59]. Under oxidizing conditions, radionuclides can react with carbonate and phosphate to form carbonate and phosphate minerals (e.g., autunite, $\text{Ca}(\text{UO}_2)_2(\text{PO}_4)_2$). Under reducing conditions, anaerobic bacteria can reduce uranium (VI) to uranium (IV) to form poorly soluble uraninite [22, 27, 35, 58]. Radium can be sorbed and co-precipitated with Fe–Mn oxyhydroxides, gypsum, barite [4, 60], and amorphous silica [61]. The stability of co-precipitated radionuclides is controlled by the host mineral solubility and stability. For example, microbial reduction of iron minerals can indirectly contribute to radium-226 release in groundwater [60].

All the processes mentioned above can act cooperatively or sequentially to control the transport behavior of radionuclide contaminants in aquifers. For example, a study by Catalano et al. [33] using the sediments collected from the Hanford 300 Area revealed that uranium co-precipitated with calcite in shallow vadoze zone sediments, formed metatorbernite ($\text{Cu}(\text{UO}_2\text{PO}_4)_2 \cdot 8\text{H}_2\text{O}$) that coexisted with adsorbed species at intermediate depths in the vadose zone, and occurred predominantly as adsorbed onto phyllosilicates in the deeper vadose zone and groundwater.

Bioreduction

Bioreduction has been proposed as a remediation approach to immobilize redox sensitive radionuclides in subsurface environments. Radionuclides can be reduced by various dissimilatory bacteria including metal-reducing bacteria and sulfate-reducing bacteria that use radionuclides as terminal electron acceptors [62, 63]. Microbial reduction of uranium (VI) to insoluble uranium (IV) by the injection of ethanol has been

demonstrated at DOE Environmental Remediation Sciences Program (ERSP) Field Research Center in Oak Ridge, Tennessee [64] and Old Rifle UMTRA field site in western Colorado. The results from Wu et al. [22] demonstrated that aqueous uranium concentrations below the USEPA maximum contaminant level ($<0.126 \mu\text{M}$) can be achieved in situ, that bioreduced/immobilized uranium is stable under anaerobic conditions, and that infiltration of dissolved oxygen into the bioreduced area promotes spatially variable oxidation of uranium (IV) and mobilization of uranium (VI). Studies at Old Rifle UMTRA site demonstrated that the immobilization of uranium (VI) in groundwater can be achieved by iron-reducing bacteria stimulated by acetate amendment at the field scale [27, 65].

However, one major unresolved question in terms of bioreduction as a viable remediation technology is the long-term stability of bioreduced radionuclides such as biogenic uranium (IV). Zhong et al. [66] demonstrated that biogenic uranium (IV) readily oxidizes once groundwater environment returns to oxic condition. Wan et al. [67] presented evidence that bioreduced uranium (IV) was reoxidized under reducing conditions because carbonate accumulation promotes the formation of highly stable carbonato-uranium (VI) complexes under neutral to slightly alkaline conditions. Further researches are needed to enable the bioreduction as a remediation technology.

Environmental Modeling Assessment

Numerical modeling can help scientists and engineers understand and predict the radioactive contaminant fate and transport in the subsurface. Specifically, the modeling can be used to integrate conceptual understanding into a consistent and numerical framework, to test hypotheses on physical and chemical processes under field relevant conditions, to plan for field experiments under uncertainties, to interpret and analyze the field experimental data, and finally to help with the design of remedial alternatives. In an iterative and complementary way, field experiments and modeling activities can work together to enable one to gain new insights and to improve the predictive capabilities on contaminant fate and transport. This section discusses two major types of approaches for modeling radionuclide fate and transport in the subsurface.

Isotherm-Based Transport Modeling

This type of models was based on the simplification that a sorption isotherm involving a single distribution coefficient (K_d) can be used to describe the sorption equilibrium of radioactive contaminants [12]. The K_d based sorption isotherm can be directly incorporated into a hydrological transport equation to simulate the coupled effects of advection, dispersion, and sorption processes [68].

As discussed before, however, the radionuclide adsorption/desorption is controlled by complex geochemical and microbiological conditions at field, and a single K_d value is usually not adequate to represent the sorption behavior over a wide range of geochemical conditions. This is because K_d values are sensitive to geochemical conditions and vary as geochemical conditions change. For example, the K_d value for uranium (VI) can vary by five orders of magnitude over the pH range from 6 to 9 and by four orders of magnitude at pH 8 as the partial pressure of CO_2 gas increases from its value in air to 0.01 atm [9, 26].

In contaminant plumes where the groundwater compositions change spatially and temporally, the K_d values could also have complex spatial patterns and evolve temporally along with the transport processes. Consequently significant errors and uncertainties may be introduced in reactive transport simulations if a constant- K_d modeling approach is used at sites where groundwater chemistry varies temporally and spatially [9, 26, 28–31].

Multi-component Reactive Transport Modeling

Over the last 2 decades, a number of models that couple advective-dispersive-diffusive transport processes with “full” geochemistry, including pH, redox-state, sediment/rock–water interactions have been developed, such as PHT3D [69], MIN3P [70], and PHAST [71]. In these coupled models, the solute transport and chemical reactions are rigorously simulated often in three-dimensional groundwater flow systems. The mechanistic treatment of chemical reactions in the coupled multi-component, multi-species mass transport has obvious advantages over the empirical isotherm-based transport models since the models more realistically account for complex geochemical processes.

In contrast to the constant- K_d modeling approach, the coupled models normally incorporate surface complexation reactions into solute transport models through the mass action equations describing the equilibria between aqueous chemical species and species formed at mineral surfaces to account for the adsorption/desorption [9]. Surface complexation models (SCM) can account for variations in chemical conditions and aqueous speciation, and thus can describe spatial and temporal changes of radionuclide adsorption [26].

There are two major approaches for applying the SCM concept to natural subsurface systems: the Component Additivity (CA) and Generalized Composite (GC) approaches [9]. The CA approach utilizes documented SCMs for well-characterized surfaces and detailed sediment characterization of the study site to determine the quantity of each reactive surface in sediments and then assembles an SCM for the sediment from its basic components [9, 26]. In the GC approach, adsorption is assumed to occur on “generic” surface sites that represent average properties of the sediment surfaces because the surface of the mineral assemblage is considered too complex to be quantified in terms of the contributions of individual phases to adsorption. Adsorption can be described by mass laws written with “generic” surface functional groups, with the stoichiometry and formation constants for each mass law determined by fitting experimental data for the mineral assemblage as a whole [9, 72].

The SCM approach has not been commonly used to describe adsorption in field-scale reactive transport modeling studies because of a poor understanding of the thermodynamics of surface complex formation in natural systems and the lack of field data. However, there is a growing application of SCM in multi-component reactive transport modeling to simulate the radionuclide fate and transport in aquifers due to recent advances in computer codes and availability of extensively geochemical characterization data in some filed sites, such as uranium reactive transport models developed at the Naturita site, Colorado [26], the Hanford site [34, 47, 73], the Oak Ridge site [35], and the Old Rifle site [27]. In addition to the models that consider equilibrium-based surface complexation reactions, models that couple kinetic mass transfer processes with surface complexation reactions have also

been developed to simulate the kinetic adsorption and desorption behavior of radionuclides. For example, Liu et al. [34, 47] proposed a multi-rate SCM to consider the effects of diffusive mass transfer on uranium (VI) adsorption/desorption processes, and the approach was evaluated at the Hanford site [73, 74].

The multi-component reactive transport model can also take into account for other biogeochemical reactions including aqueous complexation, precipitation/dissolution of radionuclide-containing minerals, and bioreduction (e.g., references [35, 75]). Thus the multi-component reactive transport model can better describe the transport and fate of radionuclide contaminants in the aquifer. However, obtaining sufficient and accurate field data for properly parameterizing a field-scale multi-component reactive transport model will remain a major challenge.

Future Directions

Future research should be aimed at improving fundamental understanding of radionuclide transport processes in heterogeneous subsurface media and facilitating transfer of knowledge and insights gained from laboratory experiment to field application.

Mass Transfer Processes of Radionuclide Contaminants in Heterogeneous Media

The variability in the physical and chemical properties of subsurface media, such as hydraulic conductivity, porosity, grain size, and reactive surface area, may vary by several orders of magnitude within an aquifer. The heterogeneity of physical and chemical properties causes spatial variations in groundwater flow velocity, reaction rate, and speciation. These variations may be associated with a range of different predominating phenomena such as preferential flow and contaminant migration pathways, hydraulically inaccessible zones into which solutes may only diffuse, and mineral grains by which solutes are selectively sorbed (e.g., references [76–82]). The individual phenomena in turn contribute to a variable extent to the spreading (dispersion) of chemical contamination, emerge and disappear at different time scales, and lead to variable reaction types and rates during solute mass transport.

As a result of physical and chemical heterogeneity, the mass transfer could occur in the subsurface at

multiple scales. The release of contaminants including cesium-137, chromium(VI), strontium-90, and uranium from contaminated sediments at the Hanford site to groundwater is found to be controlled by mass transfer in laboratory experiments [46, 55, 83–85]. Mass transfer limitations also occur at increasingly larger (macroscopic) scales in the field (e.g., references [54, 86]) such as between coarse and fine textured zones in a given facies, or between different geologic formations and facies, such as highly conductive and less conductive aquifer layers [87]. Understanding multi-scale mass transfer processes and their implications to contaminant migration and remediation at the field scale is at the forefront of reactive transport science and is a critical need for the remediation of contaminated sites [87].

Upscaling of Radionuclide Transport from Laboratory to Field Scales

Laboratory experiments provide important information on parameters and key insights for contaminant transport processes. Even complex reactive transport models may be reasonably well constrained by measured data at the laboratory scale, where a large number of measurements and observations are available. However, this is generally not the case for field-scale problems where reactive transport may be affected or controlled by strongly chemically and/or physically heterogeneous conditions (e.g., references [74, 88–91]). The heterogeneity at different scales significantly influences and drastically complicates the upscaling of solute transport and its analysis and prediction at larger scales (e.g., references [86, 92–95]). Consequently, important differences can exist between the experimental conditions under which a laboratory model was developed and calibrated, and those present in the field, including the ratios of reaction and transport time scales and variability in reactant properties and distribution.

Without appropriate consideration of the upscaling problem, reliable prediction of reactive transport processes is impossible. Therefore, scale dependence and associated upscaling of transport parameters for porous media have been actively studied for two decades. For the most part, however, the existing research has focused on the development of theoretical or empirical upscaling methods in relatively simple

systems [96–98] and on the investigation of the scale dependence of different transport parameters such as diffusion coefficients, geochemical reaction rates, sorption coefficients, and retardation factors [99] and other factors (e.g., reference [100]) at various scales ranging from pore scale to column experiments, and to field tracer tests. However, very few studies have been reported on upscaling of transport processes of radionuclide contaminants in groundwater [74]. Thus, the unsolved problem is how to systematically upscale the reactive transport process and parameters for reactive transport models involving complex, physically and chemically heterogeneous systems across multi-scales.

Bibliography

1. Bodvarsson GS, Boyle W, Patterson R, Williams D (1999) Overview of scientific investigations at Yucca Mountain – the potential repository for high-level nuclear waste. *J Contam Hydrol* 38:3–24
2. Gephart RE (2003) Hanford: a conversation about nuclear waste and cleanup. Battelle, Columbus
3. Ahearne JF (1997) Radioactive waste: the size of the problem. *Phys Today* 50(6):24–29
4. Abdelouas A (2006) Uranium mill tailings: geochemistry, mineralogy, and environmental impact. *Elements* 2:335–341
5. Riley RG, Zachara JM, Wobber FJ (1992) Chemical contaminants on DOE lands and selection of contaminant mixtures for subsurface science research, DOE/ER-0547T. U.S. Department of Energy, Washington, DC
6. Bradley DJ, Frank CW, Mikerin Y (1996) Nuclear contamination from weapons complexes in the former Soviet Union and the United States. *Phys Today* 49:40–45
7. Whicker FW, Shaw G, Voigt G, Holm E (1999) Radioactive contamination: state of the science and its application to predictive models. *Environ Pollut* 100:133–149
8. Abdelouas A, Lutze W, Nuttall HE (1999) Uranium contamination in the subsurface; characterization and remediation. *Rev Mineral Geochem* 38:433–473
9. Davis JA, Meece DE, Kohler M, Curtis GP (2004) Approaches to surface complexation modeling of uranium (VI) adsorption on aquifer sediments. *Geochim Cosmochim Acta* 68(18):3621–3641
10. Ewing RC (2004) Environmental impact of the nuclear fuel cycle. In: Gieré R, Stille P (eds) *Energy, waste and the environment: a geochemical perspective*, geological society special publication 236. The Geological Society, London, pp 7–23
11. Lee D, Walton MR, Megio JL (2005) Biological and chemical interactions with U(VI) during anaerobic enrichment in the presence of iron oxide coated quartz. *Water Res* 39:4363–4374
12. Zhu C, Anderson G (2002) *Environmental applications of geochemical modeling*. Cambridge University Press, London
13. Renshaw J, Butchins LJC, Livens FR, May I, Charnock JM, Lloyd JR (2005) Bioreduction of uranium: environmental implications of a pentavalent intermediate. *Environ Sci Technol* 39:5657–5660
14. Fishlock D (1994) The dirtiest place on earth. *New Sci* 1913:34–37
15. Zachara JM, Serne J, Freshley M, Mann F, Anderson F, Wood M, Jones T, Myers D (2007) Geochemical processes controlling migration of tank wastes in Hanford's vadose zone. *Vadose Zone J* 6:985–1003
16. Gee GM, Oostrom M, Freshley MD, Rockhold ML, Zachara JM (2007) Hanford site vadose zone studies: an overview. *Vadose Zone J* 6:899–905
17. Um W, Serne RJ, Brown CF, Last GV (2007) U(VI) adsorption on aquifer sediments at the Hanford site. *J Contam Hydrol* 93:255–269
18. Hartman MJ, Morasch LF, Webber WD (2007) Hanford site groundwater monitoring for fiscal year 2006, Pacific Northwest National Laboratory, Richland
19. Hartman MJ, Webber WD, Fluor Hanford Inc (2008) Hanford site groundwater monitoring for fiscal year 2007. DOE/RL-2008-01, Revision 0, Pacific Northwest National Laboratory, Richland
20. Phillips H, Watson DB, Roh Y (2007) Uranium deposition in a weathered fractured saprolite/shale. *Environ Sci Technol* 41:7653–7660
21. Wu W, Carley J, Fienen M, Mehlhorn T, Lowe K, Nyman J, Luo J, Gentile ME, Rajan R, Wagner D, Hickey RF, Gu B, Watson D, Cirpka O, Kitanidis P, Jardine J, Criddle CS (2006) Pilot-scale in situ bioremediation of uranium in a highly contaminated aquifer. 1. Conditioning of a treatment zone. *Environ Sci Technol* 40(12):3978–3985
22. Wu W, Carley J, Luo J, Ginder-Vogel MA, Cardenas E, Leigh MB, Hwang C, Kelly SD, Ruan C, Wu L, Nostrand JV, Gentry T, Lowe K, Mehlhorn TL, Carroll S, Luo W, Fields MW, Gu B, Watson D, Kemner K, Marsh T, Tiedje J, Zhou J, Fendorf S, Kitanidis PK, Jardine PM, Criddle C (2007) In situ bioreduction of uranium (VI) to submicromolar levels and reoxidation by dissolved oxygen. *Environ Sci Technol* 41:5716–5723
23. Carlton WH (1997) Assessment of neptunium, americium, and curium in the Savannah River Site Environment. WSRC-TR-97-00266, Westinghouse Savannah River Co, Aiken
24. Dai M, Kelley JM, Buesseler KO (2002) Sources and migration of plutonium in groundwater at the Savannah River Site. *Environ Sci Technol* 36:3690–3699
25. Westinghouse Savannah River Co (1998) The Savannah River Site's groundwater monitoring program: third quarter 1997. ESH-EMS-970490, U.S. Department of Energy, Washington, DC
26. Curtis GP, Davis JA, Naftz DL (2006) Simulation of reactive transport of uranium (VI) in groundwater with variable chemical conditions. *Water Resour Res* 42:W04404. doi:10.1029/2005WR003979
27. Yabusaki SB, Fang Y, Long PE, Resch CT, Peacock AD, Komlos J, Jaffed PR, Morrison SJ, Dayvault RD, White DC, Anderson RT (2007) Uranium removal from groundwater via in situ biostimulation: field-scale modeling of transport and biological processes. *J Contam Hydrol* 93:216–235

28. Read D, Ross D, Sims RJ (1998) The migration of uranium through Clashach sandstone: the role of low molecular weight organics in enhancing radionuclide transport. *J Contam Hydrol* 35:235–248
29. Bethke CM, Brady PV (2000) How the Kd approach undermines group water cleanup. *Ground Water* 38(3):435–443
30. Glynn PD (2003) Modeling Np and Pu transport with a surface complexation model and spatially variant sorption capacities; implications for reactive transport modeling and performance assessments of nuclear waste disposal sites; reactive transport modeling in the geosciences. *Comput Geosci* 29(3):331–349
31. Zhu C (2003) A case against Kd-based transport models: natural attenuation at a mill tailings site; reactive transport modeling in the geosciences. *Comput Geosci* 29(3):351–359
32. Bond DL, Davis JA, Zachara JM (2008) Uranium(VI) release from contaminated vadose zone sediments: estimation of potential contributions from dissolution and desorption. In: Barnett MO, Kent DB (eds) *Adsorption of metals to geomedia II*. chap. 14, Elsevier, Amsterdam, pp 375–416
33. Catalano JG, McKinley JP, Zachara JM, Heald SM, Smith SC, Brown GE (2006) Changes in uranium speciation through a depth sequence of contaminated Hanford sediments. *Environ Sci Technol* 40(8):2517–2524
34. Liu C, Zachara JM, Qafoku NP, Wang Z (2008) Scale-dependent desorption of uranium from contaminated subsurface sediments. *Water Resour Res* 44:W08413. doi:10.1029/2007WR006478
35. Luo J, Weber F, Cirpka OA, Wu W, Nyman JL, Carley J, Jardine PM, Criddle CS, Kitanidis PK (2007) Modeling in-situ uranium (VI) bioreduction by sulfate-reducing bacteria. *J Contam Hydrol* 92:129–148
36. Keeney-Kennicutt WL, Morse JW (1985) The redox chemistry of Pu(V)O_2^+ interaction with common mineral surfaces in dilute solutions and seawater. *Geochim Cosmochim Acta* 49(12):2577–2588
37. Sanchez AL, Murray JW, Sibley TH (1985) The adsorption of plutonium on goethite. *Geochim Cosmochim Acta* 49:2297–2307
38. Duff MC, Hunter DB, Triay IR, Bertsch PM, Reed DT, Sutton SR, Shea-McCarthy G, Kitten J, Eng P, Chipera SJ, Vaniman DT (1999) Mineral associations and average oxidation states of sorbed Pu on tuff. *Environ Sci Technol* 33:2163–2169
39. Novikov P, Kalmykov SN, Utsunomiya S, Ewing RC, Horreard F, Merkulov A, Clark SB, Tkachev VV, Myasoedov BF (2006) Colloid transport of plutonium in the far-field of the mayak production association, Russia. *Science* 314:638–641
40. Kim JI (1993) The chemical behavior of transuranium elements and barrier functions in natural aquifer systems. *Mater Res Soc Symp Proc* 294:3–21
41. Kim JI (1994) Actinide colloids in natural aquifer systems. *Mater Res Soc Bull* 19:47–53
42. Braithwaite A, Livens FR, Richardson S, Howe MT (1997) Kinetically controlled release of uranium from soils. *Eur J Soil Sci* 48:661–673
43. Barnett MO, Jardine PM, Brooks SC, Selim HM (2000) Adsorption and transport of uranium(VI) in subsurface media. *Soil Sci Soc Am J* 64:908–917
44. Baik MH, Cho WJ, Hahn PS (2004) Sorption of U(VI) onto granite surfaces: a kinetic approach. *J Radioanal Nucl Chem* 260:495–502
45. Qafoku NP, Zachara JM, Liu C, Gassman PL, Qafoku OS, Smith SC (2005) Kinetic desorption and sorption of U(VI) during reactive transport in a contaminated Hanford sediment. *Environ Sci Technol* 39:3157–3165
46. Liu C, Zachara JM, Yantasee W, Majors PD, McKinley JP (2006) Microscopic reactive diffusion of uranium in the contaminated sediments at Hanford, USA. *Water Resour Res* 42:W12420. doi:10.1029/2006WR005031
47. Liu C, Shi S, Zachara JM (2009) Kinetics of uranium(VI) desorption from contaminated sediments: effect of geochemical conditions and model evaluation. *Environ Sci Technol* 43(17):6560–6566
48. Arai Y, Marcus MA, Tamura N, Davis JA, Zachara JM (2007) Spectroscopic evidence for uranium bearing precipitates in vadose zone sediments at the Hanford 300-area site. *Environ Sci Technol* 41:4633–4639
49. Stubbs JE, Veblen LA, Elbert DC, Zachara JM, Davis JA, Veblen DR (2009) Newly recognized hosts for uranium in the Hanford site vadose zone. *Geochim Cosmochim Acta* 73(6):1563–1576
50. Waite TD, Davis JA, Payne TE, Waychunas GA, Xu N (1994) Uranium (VI) adsorption to ferrihydrite: application of a surface complexation model. *Geochim Cosmochim Acta* 58(24):5465–5478
51. Fox PM, Davis JA, Zachara JM (2006) The effect of calcium on aqueous uranium(VI) speciation and adsorption to ferrihydrite and quartz. *Geochim Cosmochim Acta* 70:1379–1387
52. Dong W, Ball WP, Liu C, Wang Z, Stone AT, Bai J, Zachara JM (2005) Influence of calcite and dissolved calcium on uranium (VI) sorption to a Hanford subsurface sediment. *Environ Sci Technol* 39:7949–7955
53. Hu QH, Zavarin M, Rose TP (2008) Effect of reducing groundwater on the retardation of redox-sensitive radionuclides. *Geochem Trans* 9:12. doi:10.1186/1467-4866-9-12
54. Liu C, Zachara JM, Qafoku OS, McKinley JP, Heald SM, Wang Z (2004) Dissolution of uranyl microprecipitates in subsurface sediments at Hanford site, WA. *Geochim Cosmochim Acta* 68:4519–4537
55. McKinley JP, Zachara JM, Liu C, Heald SM (2006) Microscale controls on the fate of contaminant uranium in the vadose zone, Hanford site, Washington. *Geochim Cosmochim Acta* 70:1873–1887
56. Gómez P, Garraón A, Buil B, Turrero MJ, Sánchez L, de la Cruz B (2006) Modeling of geochemical processes related to uranium mobilization in the groundwater of a uranium mine. *Sci Total Environ* 366:295–309
57. Abdelouas A, Lutze W, Nuttall E (1998) Chemical reactions of uranium in ground water at a mill tailings site. *J Contam Hydrol* 34:343–361

58. Abdelouas A, Lutze W, Gong W, Nuttall EH, Strietelmeier BA, Travis BJ (2000) Biological reduction of uranium in groundwater and subsurface soil. *Sci Total Environ* 250:21–35
59. Ohnuki T, Kozai N, Samadfam M, Yasuda R, Yamamoto S, Narumi K, Naramoto H, Murakami T (2004) The formation of autunite ($\text{Ca}(\text{UO}_2)_2(\text{PO}_4)_2 \cdot n\text{H}_2\text{O}$) within the leached layer of dissolving apatite: incorporation mechanism of uranium by apatite. *Chem Geol* 211:1–14
60. Martin AJ, Crusius J, Jay McNee J, Yanful EK (2003) The mobility of radium-226 and trace metals in pre-oxidized subaqueous uranium mill tailings. *Appl Geochem* 18:1095–1110
61. Landa ER (2004) Uranium mill tailings: nuclear waste and natural laboratory for geochemical and radioecological investigations. *J Environ Radioactiv* 77:1–27
62. Lovely DR, Coates JD (1997) Bioremediation of metal contamination. *Curr Opin Biotechnol* 8:285–289
63. Liu C, Gorby YA, Zachara JM, Fredrickson JK, Brown CF (2002) Reduction Kinetics of Fe(III), Co(III), U(VI), Cr(VI), and Tc (VII) in cultures of dissimilatory metal-reducing bacteria. *Biotechnol Bioeng* 80(6):637–649
64. Wu W, Carley J, Gentry T, Ginder-Vogel MA, Fienen M, Mehlhorn T, Yan H, Carroll S, Nyman J, Luo J, Gentile ME, Fields MW, Hickey RF, Watson D, Cirpka OA, Fendorf S, Zhou J, Kitanidis P, Jardine PM, Criddle CS (2006) Pilot-scale in situ bioremediation of uranium in a highly contaminated aquifer. 2: U(VI) reduction and geochemical control of U(VI) bioavailability. *Environ Sci Technol* 40:3986–3995
65. Anderson RT, Vrionis HA, Ortiz-Bernad I, Resch CT, Long PE, Dayvault R, Karp K, Marutzky S, Metzler DR, Peacock A, White DC, Lowe M, Lovley DR (2003) Stimulating the in situ activity of *Geobacter* species to remove uranium from the groundwater of a uranium-contaminated aquifer. *Appl Environ Microbiol* 69(10):5884–5891
66. Zhong L, Liu C, Zachara JM, Kennedy DW, Szecsody JE, Wood BD (2005) Oxidative remobilization of biogenic uranium (IV) precipitates: effects of Iron (II) and pH. *J Environ Qual* 34(5):1763–1771
67. Wan J, Tokunaga TK, Brodie E, Wang Z, Zheng Z, Herman D, Hazen T, Firestone MK, Sutton SR (2005) Reoxidation of bio-reduced uranium under reducing conditions. *Environ Sci Technol* 39:6162–6169
68. Zheng C, Wang PP (1999) MT3DMS, a modular three-dimensional multi-species transport model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems; documentation and user's guide, U.S. Army Engineer Research and Development Center Contract Report SERDP-99-1, Vicksburg, p 202. <http://hydro.geo.ua.edu/mt3d/>
69. Prommer H, Barry DA, Zheng C (2003) MODFLOW/MT3DMS based reactive multicomponent transport modeling. *Ground Water* 41(2):247–257
70. Mayer KU, Frind EO, Blowes DW (2002) Multicomponent reactive transport modeling in variably saturated porous media using a generalized formulation for kinetically controlled reactions. *Water Resour Res* 38:1174. doi:10.1029/2001WR000862
71. Parkhurst DL, Kipp KL, Engesgaard P, Charlton SC (2004) PHAST – a program for simulating ground-water flow, solute transport and multicomponent geochemical reactions. USGS Tech Meth 6-A8:154
72. Davis JA, Payne TE, Waite TD (2002) Simulating the pH and pCO_2 dependence of uranium(VI) adsorption by a weathered schist with surface complexation models. In: *Geochemistry of soil radionuclides*. Soil Science Society of America, Madison, pp 61–68
73. Ma R, Zheng C, Prommer H, Greskowiak J, Liu C, Zachara J, Rockhold M (2010) A field-scale reactive transport model for U (VI) migration influenced by coupled multirate mass transfer and surface complexation reactions. *Water Resour Res* 46: W05509. doi:10.1029/2009WR008168
74. Greskowiak J, Prommer H, Liu C, Post VEA, Ma R, Zheng C, Zachara JM (2010) Comparison of parameter sensitivities between a laboratory and field scale model of uranium transport in a dual domain, distributed-rate reactive system. *Water Resour Res* 46:W09509. doi:10.1029/2009WR008781
75. Fang Y, Yabusaki SB, Morrison SJ, Amonette JP, Long PE (2009) Multicomponent reactive transport modeling of uranium bioremediation field experiments. *Geochim Cosmochim Acta* 73:6029–6051
76. Gelhar LW (1986) Stochastic subsurface hydrology – from theory to applications. *Water Resour Res* 22(9):1355–1455
77. Dagan G (1989) Flow and transport in porous formations. Springer, New York
78. Barber LB (1994) Sorption of chlorobenzenes to Cape Cod aquifer sediments. *Environ Sci Technol* 28:890–897
79. Friedly JC, Davis JA, Kent DB (1995) Modeling hexavalent chromium reduction in groundwater in field-scale transport and laboratory batch experiments. *Water Resour Res* 31:2783–2794
80. Kleineidam S, Rugner H, Grathwohl P (1999) Impact of grain scale heterogeneity on slow sorption kinetics. *Environ Toxicol Chem* 18:1673–1678
81. Allen-King RM, Divine DP, Robin MJL, Alldredge JRG (2006) Spatial distributions of perchloroethylene reactive transport parameters in the Borden aquifer. *Water Resour Res* 42:W01413. doi:10.1029/2005WR003977
82. Descourvières C, Hartog N, Patterson BM, Oldham C, Prommer H (2010) Geochemical controls on sediment reactivity and buffering processes in a heterogeneous aquifer. *Appl Geochem* 25:261–275
83. Liu C, Zachara JM, Smith SC, McKinley JP, Ainsworth CC (2003) Desorption kinetics of radiocesium from the subsurface sediments at Hanford site, USA. *Geochim Cosmochim Acta* 67:2893–2912
84. McKinley JP, Zachara JM, Smith SC, Liu C (2007) Cation exchange reactions controlling desorption of $^{90}\text{Sr}^{2+}$ from coarse-grained contaminated sediments from the Hanford formation, Washington. *Geochim Cosmochim Acta* 71(2):305–325
85. Zachara JM, Ainsworth CC, Brown GE Jr, Catalano JG, McKinley JP, Qafoku O, Smith SC, Szecsody JE, Traina SJ, Warner JA (2004)

- Chromium speciation and mobility in a high level nuclear waste vadose zone plume. *Geochim Cosmochim Acta* 68(1):13–30
86. Zheng C, Gorelick SM (2003) Analysis of solute transport in flow fields influenced by preferential flow paths at the decimeter scale. *Ground Water* 41(2):142–155
 87. Zachara J, Freshley M, Andersen G, DePaolo D, Fredrickson J, Haggerty R, Kent D, Konopka A, Lichtner P, Liu C, McKinley J, Rockhold M, Rubin Y, Szecsody J, Versteeg R, Ward A, Williams B, Zheng C (2007) Integrated field-scale subsurface research challenge, multi-scale mass transfer processes controlling natural attenuation and engineered remediation: an IFC focused on Hanford's 300 area uranium plume. Proposal to the U.S. Department of Energy Office of Biological and Environmental Research LAB 06–16 – Environmental Remediation Science Program
 88. Morrison SJ, Tripathi VS, Spangler RR (1995) Coupled reaction/transport of a chemical barrier for controlling U(VI) contamination in groundwater. *J Contam Hydrol* 17:347–363
 89. Zhu C, Hu FQ, Burden DS (2001) Multi-component reactive transport modeling of natural attenuation of an acid groundwater plume at a uranium mill tailings site. *J Contam Hydrol* 52:85–108
 90. Bain JG, Mayer KU, Blowes DW, Frind EO, Molson JWH, Kahnt R, Jenk U (2001) Modeling the closure-related geochemical evolution of groundwater at a former uranium mine. *J Contam Hydrol* 52:109–135
 91. Yabusaki SB, Fang Y, Waichler SR (2008) Building conceptual models of field-scale uranium reactive transport in a dynamic vadose zone-aquifer-river system. *Water Resour Res* 44: W12403. doi:10.1029/2007WR006617
 92. Feehley CE, Zheng C, Molz FJ (2000) A dual-domain mass transfer approach for modeling solute mass transfer in heterogeneous porous media, application to the MADE site. *Water Resour Res* 36:2501–2515
 93. Haggerty R, Harvey CF, von Schwerin CF, Meigs LC (2004) What controls the apparent timescale of solute mass transfer in aquifers and soils? A comparison of experimental results. *Water Resour Res* 40:W01510. doi:10.1029/2002WR001716
 94. Gorelick SM, Liu G, Zheng C (2005) Quantifying mass transfer in permeable media containing conductive dendritic networks. *Geophys Res Lett* 32:L18402. doi:10.1029/2005GL023512
 95. Zheng C, Bianchi M, Gorelick SM (2010) Lessons learned from 25 years of research at the MADE site. *Ground Water* (in press)
 96. Seebonruang U, Ginn TR (2006) Upscaling heterogeneity in aquifer reactivity via exposure-time concept: forward model. *J Contam Hydrol* 84:127–154
 97. Fernández-García D, Llerar-Meza G, Gómez-Hernández JJ (2009) Upscaling transport with mass transfer models: mean behavior and propagation of uncertainty. *Water Resour Res* 45:W10411. doi:10.1029/2009WR007764
 98. Heße F, Radu FA, Thullner M, Attinger S (2009) Upscaling of the advection–diffusion–reaction equation with monod reaction. *Adv Water Resour* 32:1336–1351
 99. Deng H, Dai Z, Wolfsberg A, Lu Z, Ye M, Reimus P (2010) Upscaling of reactive mass transport in fractured rocks with multimodal reactive mineral facies. *Water Resour Res* 46:15. doi:10.1029/2009WR008363
 100. Wang F, Bright J (2004) Scale effect and calibration of contaminant transport models. *Ground Water* 42(5):760–766

Groundwater Remediation, Environmental and Economic Assessment

PAUL HARDISTY¹, ECE OZDEMIROGLU²

¹WorleyParsons, Perth, WA, Australia

²Economics for the Environment Consultancy (EFTEC), London, UK

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Decision-Making Context
The Economics of Remediation
Case Histories
Discussion and Conclusions
Future Directions
Bibliography

Glossary

CBA Cost benefit analysis – A decision-making tool that compares costs and benefits of a proposed policy or project in monetary terms. The aim is to find the most beneficial (highest net benefit) option of achieving a given objective. Taking no action is also included as an option.

CEA Cost effectiveness analysis – A decision-making tool that compares the costs of a proposed policy or project to a (nonmonetary) measure of its benefits. The aim is to find the option that generates the highest benefit for each unit of money spent.

TEV Total economic value – The economic value of a resource comprised of its use and nonuse values.

Definition of the Subject and Its Importance

As the world's population grows, industrial activities continue to degrade land and water at a faster pace. Under the “polluter pays” principle, increasingly adopted as the fundamental ethical precept for remediation policy, the responsibility for planning, funding, and executing remediation lies with the polluter. And ever present and increasingly vocal and powerful, are the public, the neighbors, the inhabitants of the planet, demanding that their interests be served also, and that the planet's dwindling resources be protected for their future, and the future of their children.

But remediating polluted aquifers can be expensive, technically difficult, and time-consuming [1, 2]. Deciding if and when to remediate, and to what degree, can be regarded in the context of alternative environmentally and socially beneficial actions. What else could be done with the money required to restore a site or aquifer? Which of the many choices available would provide the greatest benefit to society? And then, what are the commercial realities facing those who are called upon to pay for remediation?

Significant amounts of time, effort, and money have already been devoted to remediation of aquifers worldwide. A tremendous diversity of methods and technologies has been applied, in conditions as variable as the individual sites themselves. Along the way, consultants, problem holders, individual professionals, and government institutions have accumulated wide knowledge of the costs of remediation.

Combining and prioritizing the diverse objectives, interests, and constraints of different stakeholders involved requires a multidisciplinary approach to decision-making. Requirements include understanding the objectives, the risks (of taking remediation action and no action), and the environmental and financial costs and benefits of remediation. Combining information and data from different disciplines requires a common unit of value for direct comparison.

Truly sustainable remediation requires moving beyond money and examining which of a variety of options best optimizes overall value to society over the longer term. Given the inherent uncertainty associated with valuing market and nonmarket goods, even in the present, the relative overall long-term performance of

remedial options must be evaluated over a wide range of possible future conditions.

Introduction

Until very recently, selecting the least-cost remedial option passed for “economic” analysis [3, 4]. The benefits to the problem holder were sometimes considered; the wider benefits to other parts of society rarely [5]. Borrowing from the wider environmental economics literature [6], the costs and wider economic benefits of remedial alternatives can be compared to select optimal remediation approaches [7]. A critical part of this equation, rarely considered, is the cost of secondary effects or by-products of the remedial action, including the so-called external costs of many common remedial practices, such as excavation and land-filling contaminated soils and materials (discussed below). This entry presents the case for applying an environmental and economic multidisciplinary approach to assessing the costs and benefits of groundwater remediation. This combined approach is termed environmental and economic sustainability assessment (EESA) to remedial decision-making [8]. Environmental approaches within the EESA are used to identify the risks created by contamination and the ability of remediation options to address these risks. Economic approaches are used to quantify the cost to the receptors (the environment itself, users of the environment, water users, and so on) of contamination.

Decision-Making Context

At the outset, it is important to distinguish between the different levels at which remedial decisions need to be made. The current literature makes reference to “remedial approaches,” “remedial options,” and “remedial technologies,” sometimes interchangeably, and often without clear definition. For contaminated groundwater (and indeed for contaminated land sites), the distinctions between objectives, approaches (or strategies) and technologies is vitally important. These are formally defined below [1].

- *Remedial objective* is the overall intent of the remediation. Objectives could include the degree to which groundwater is to be remediated, the

protection of specific receptors, or the elimination or reduction of certain unacceptable risks.

- *Remedial strategy* (or approach) is the way in which the remediation objective is to be reached, and is defined specifically in terms of the pollutant linkage component it addresses: source removal, pathway elimination, source protection/isolation, or a combination thereof.
- *Remedial technologies* are the specific tools which form the components of the remediation strategy. For example, physical containment (a pathway elimination approach) can be achieved through use of slurry walls, sheet pile walls, or liners, often in conjunction with groundwater pumping and treatment. Source removal can be achieved through excavation and on-site treatment of contaminated soils (by a variety of techniques), or through many available in situ mass destruction techniques. A remedial strategy will very often involve the use of several different remedial technologies.

These levels are all interlinked. The remedial objective should be known before detailed design (technology selection) occurs. Designing options to meet the same strategic objective is crucial to ensure a consistent comparison of costs and benefits. The choice of a remedial approach is a critical intermediate step, which can be used both to help set objectives (by considering and comparing various approaches at the conceptual level), and to guide the selection of the technological components that will make up the final design. Each of these three levels of analysis is discussed below.

Contamination issues must be seen in the context of time and space, and are inherently dynamic in nature. This presents a number of challenges for the setting of remedial objectives and assessing the most economic remediation alternative: (1) objectives must be framed in a temporal context; (2) technology changes with time, and (3) regulations change with time. In the same way, the *scale* of a contamination problem is not necessarily fixed. A spill which is initially concentrated in a small area may over time expand and affect a considerable volume, as contaminants migrate laterally and vertically, bringing them into contact with other media and

receptors. The scale of a contamination problem may have significant impact on how it is valued by society.

Remedial Objective

The remedial objective is the level at which the benefits of remediation are most readily and fundamentally determined. If a valuable receptor is protected, a benefit to society accrues. If a receptor is not protected, damage (and cost to society) results. Benefits are tied clearly to the fundamental remedial objective, and the basic approach used to achieve it. For a groundwater contamination problem, the choice of whether to achieve the objective using pump-and-treat, a bio-barrier, or natural attenuation, has a direct impact on costs (including any disbenefits associated with the method such as release of off-gases to the atmosphere, for instance), but benefits remain essentially constant.

Choosing a remedial objective can become quite complex when mobile groundwater plumes are involved. Consider, for example, a fixed point source which actively introduces contaminants into groundwater at a mass rate (dm/dt). A conservative solute, such as chloride, will move at the linear advective groundwater velocity. Many organic contaminants, such as benzene for example, will biologically degrade over time, and are also subject to adsorption onto matrix material. As time passes, the plume migrates in groundwater, dispersing laterally and transversely due to the effects of chemical diffusion and mechanical mixing. Early in the plume's evolution, the contaminants may only have migrated a short distance, and might still be quite highly concentrated. Only a relatively small volume of aquifer has been impacted. At this stage, perhaps no receptors of concern have yet been impacted. But as plume migration continues unabated, receptors may become impacted. The further the plume spreads, the greater the likelihood that more distant receptors will be affected. The result is that the number of receptors impacted may increase with time. Depending on the behavior of the contaminant solute (degree of attenuation by adsorption, dispersion, biological and chemical degradation), impacts could also vary with time at a given receptor. In such a situation, risk needs to be described as a function of time and space. As the plume migrates and disperses

with time, various receptors are impacted at different times. For interventions which take place at a given point in time and space ($[x, t]$ coordinates), a remedial cost and benefit (equal to damage avoided if the remediation took place) would be realized. So for the generic case, costs and benefits of remediation vary not only with time, but with the location in space at which the remedial action is implemented. The various remedial objective options can be evaluated within this context.

Remedial objectives might include preventing or reducing the impact to a specific receptor, or preventing contamination from exceeding a given concentration at a specified point. If a receptor has already been impacted by the time the problem is discovered, the objective may be to prevent further exposure, or even attempt reversal of the damage. Remedial objectives may also change with time. An initial evaluation could indicate that net benefits are maximized if a certain noncritical receptor could be sacrificed, and that the situation with respect to a more distant receptor could be reevaluated, given that migration times would be expected to be long and attenuation active. At the reevaluation point, remedial technology options may have broadened, the value of the receptor may have increased, and the degree of attenuation and migration may have turned out to be different from that which was originally predicted. A reanalysis of the costs and benefits could indicate that a change of objectives is warranted. Flexibility based on continuous monitoring of remedial progress is an important part of any remedial strategy.

Remedial Strategy

There exist a staggering number of different remediation technologies available to achieve any specific technical outcome. For example, dissolved volatile organic compounds can be removed from the subsurface by physically pumping out the water and treating it at surface with a variety of methods, can be destroyed in situ by promoting aerobic or anaerobic biodegradation, can be volatilized in situ by introducing air into the saturated zone (air sparging), or can be removed passively with a variety of permeable barrier systems. Available technologies for controlling the migration of contaminants in groundwater include physical barrier systems,

funnel-and-gate technology, hydraulic control through pumping and reinjection, and permeable reactive barriers. Sources of ongoing groundwater contamination (often called subsurface sources), including NAPLs (non-aqueous phase liquids), can be removed in a wide variety of ways. Soil remediation is often a critical part of an overall groundwater remediation program. Contaminants can be removed from soil by excavation and treatment using methods such as soil washing, enhanced biological treatment, or chemical oxidation, or with a variety of in situ methods. Often, several different technologies, each designed to achieve a specific technical outcome, will be required to accomplish the remedial objective. Because so many technologies exist, which achieve such different technical outcomes, at such widely varying costs, direct comparison of technologies using EESA is often not practical or useful, without first framing them within a set of remedial strategies [9].

The remedial strategy does not focus on technology per-se, but on ways of breaking the pollutant risk linkage, which causes (or will cause) damage. The list of possible remedial strategies is relatively short: either remove the source, eliminate the pathway, or protect/move/manage the receptor. Consideration of remedial strategy can be very useful in aligning the EESA with current risk-based guidance, and in streamlining the EESA process, since only a limited number of strategies need be considered [8]. Remedial strategy provides a link between remedial objectives, and the hundreds of clean-up technologies available. Also, the degree to which the risk linkage is broken, the timing of the action, and the spatial location at which the action is taken, are all variables which must be considered when choosing a strategy. Constraints analysis can be undertaken to help assess which strategies are realistically achievable. Preliminary high-level costs can be assigned to each strategy which can feasibly achieve the desired objective, and compared to the benefits of achieving the objective. This provides a relatively quick strategic analysis of the costs and benefits of remediation, and a basis for selection of an optimal remedial strategy, before proceeding to detailed technology evaluation and cost analysis.

Remedial Technology

The remedial technology selection level involves choosing the most cost-effective way of implementing

a remedial strategy. This requires detailed comparison of capital and operation and maintenance (O&M) costs for technologies over a projected project life span. The external costs of remediation over the life cycle would also be incorporated into the cost analysis. Application of constraints to remediation helps to reduce the number of viable technologies which can be feasibly applied to the problem. In some cases, different technologies may also realize different benefits (within a given strategy). If these are significant, they should be included in the assessment.

The Economics of Remediation

Introduction to Environmental Economic Analysis

Environmental resources, groundwater included, have economic values that are determined by individuals' preferences for them. Economic values are expressed in relative terms based on individuals' preferences for given *changes* in the quality and/or quantity of resources. The unit used for economic valuation is money – as it is a common unit making the comparison of financial and environmental costs and benefits possible.

Many environmental resources or services they provide are market goods and services. In the context of groundwater, the key market good is water abstracted for public water supply and other uses. The market price at which a good is exchanged reveals some information on its economic value. In particular, for the buyer of a good, the price reveals the amount of money the buyer is at least willing to give up to obtain the good. For the seller, the price reveals the amount of money the seller is at least willing to accept as compensation for giving up the good. Using this unit, preferences are measured in terms of individuals' willingness to pay (WTP) money to avoid an environmental loss or to secure a gain and their willingness to accept (WTA) money as compensation to tolerate an environmental loss or to forgo a gain.

Many resources and services supported by environmental resources are not traded in markets and are consequently “unpriced” or “nonmarket” goods and services. Most groundwater services such as recharging surface waters are nonmarket. These nonmarket goods and services still have economic values.

People have several motivations for having positive WTP and WTA, which are presented in the Total

Economic Value (TEV) typology. Total Economic Value is the sum of:

Use values, which involve some interaction with the resource, either directly or indirectly:

- *Direct use value*: The use of groundwater in either a consumptive manner, such as industrial water abstraction or in a nonconsumptive manner such as for recreation (e.g., fishing) from the surface waters supported by groundwater.
- *Indirect use value*: The role of groundwater in providing or supporting key (ecosystem) services, such as nutrient cycling, habitat provision, water regulation and cleaning, etc.
- *Option value*: Not associated with current use of groundwater but the benefit of keeping open the option to make use of groundwater in the future. A related concept is *quasi-option value*, which arises through avoiding or delaying irreversible decisions, where technological and knowledge improvements can alter the optimal management of a natural resource.

Nonuse (or passive use) value is associated with benefits derived simply from the knowledge that the natural resources and aspects of the natural environment are maintained, that is, it is not associated with any use of a resource. For example, individuals place a value to knowing that aquifers will be protected from pollution or over-abstraction even though they have no intention to make any direct or indirect use. Nonuse value can be split into three parts:

- *Altruistic value*: Derived from knowing that contemporaries can enjoy the goods and services provided by groundwater
- *Bequest value*: Associated with the knowledge that groundwater as a resource will be passed on to future generations
- *Existence value*: Derived simply from the satisfaction of knowing that aquifers continue to exist, regardless of use made of it by oneself or others now or in the future

Those who make direct and indirect use of an environmental resource, that is, the users, are likely to hold both use and nonuse values. Those who do not directly or indirectly use the resource but still hold nonuse values are called nonusers. While users are

relatively easy to identify (e.g., owners of the site to be remediated or owners of surrounding properties – see below regarding blight), there is no theoretical definition of nonusers. The definition is an empirical question which can be answered by primary research.

Two types of economic valuation methods are developed to estimate the economic value of the nonmarket goods and services in the absence of price information. The first type is called revealed preference methods. They use price and consumption information from markets that are affected by environmental quality. For example, hedonic property pricing method estimates the premium buyers pay for properties in environmentally high-quality surroundings. Travel cost estimates the economic value of informal (free of direct charge) recreation by analyzing the costs incurred by recreational visitors to travel to and from and at a recreational site. The second type is called stated preference methods, which use questionnaires to elicit individuals' WTP and/or WTA. These methods are potentially applicable to any resource and decision context and the only methods that can estimate nonuse values.

Economic value evidence is used in cost benefit analysis (CBA) as well as other decision-making contexts. CBA is a decision framework that compares the costs and benefits of a policy or project action and nonaction. The costs and benefits have precise meanings. Any change that is likely to increase society's welfare generates a benefit. Any change that is likely to decrease society's welfare generates a cost. An avoided cost (e.g., cost of contamination avoided through remediation) becomes a benefit. Similarly, a foregone benefit (e.g., if remediation does not take place) becomes a cost.

Once each type of cost and benefit is estimated in monetary terms, they are aggregated over different types of costs and benefits, over affected populations, and over time. The comparison of these aggregated costs and benefits is done through two calculations. The first is the Net Present Value (NPV), which is the sum of net benefits (benefits minus costs) each year over time. As the value of a benefit (or a cost) is less in the future for various reasons, the future costs and benefits are discounted. A positive (negative) NPV means that aggregate benefits over time exceed (are less than) aggregate costs over time. CBA recommends

options with positive NPV for further consideration. The option with the highest positive NPV is the best option according to the NPV calculation. The second comparison calculation is the Benefit Cost Ratio (BCR), which is the ratio of benefits summed over time (and discounted) to costs summed over time (and discounted). A BCR greater (less) than 1 means that aggregate benefits over time exceed (are less than) aggregate costs over time. CBA recommends options with BCR greater than 1. The option with the highest BCR is the best option according to the NPV calculation. NPV is the preferred calculation when the decision-making context is one of finding the option with the highest positive welfare impact. BCR is the preferred calculation when the decision-making context is one of finding the best (highest welfare) way of spending a given budget.

CBA allows for all options, including not taking action, to be compared so long as all options are set to achieve the same objective. Cost Effectiveness Analysis (CEA), on the other hand, estimates the costs in the same way as CBA but does not quantify the benefits in monetary term. The CEA ratio is then like BCR where benefits are in some physical or other quantity as relevant to the decision-making context. The shortcoming of CEA compared to CBA is that the former cannot judge the cost-effectiveness of taking no action and hence whether it is worthwhile to achieve the given objective at all.

CBA requires a multidisciplinary approach in order to identify and quantify the costs and benefits before their economic (monetary) value can be estimated. Even an extensive CBA is not always able to quantify and monetize all costs and benefits due to lack of data and uncertainties. Therefore, sensitivity analysis for testing different values and parameters is crucial.

Such a multidisciplinary approach has been developed in the context of groundwater remediation. The EESA (environmental and economic sustainability assessment) method employs elements of CBA, risk assessment, and sensitivity analysis, along with a full technical analysis of remedial options, effectiveness, and secondary impacts, to assist in identifying remediation strategies and options, which will result in optimal net benefit to all stakeholders. At the end of the process, it is about comparing possible courses of action, using money as a common unit of measure to

examine necessary trade-offs, but then moving beyond money to examine which options are most robustly superior under the widest range of possible future conditions.

What types of costs and benefits and what rate of discounting applies are discussed below in the specific context of groundwater. Case histories illustrate how to use NPV and BCR.

Economic Analysis in the Context of Groundwater Remediation

An economic assessment framework was specifically developed for groundwater remediation cases in Hardisty and Ozdemiroglu [7]. In this analysis, it is assumed that there is full knowledge that the contamination has occurred, and that damage is occurring. Situations where damage is occurring without the knowledge of the public or regulators are not considered here, though the same method is applicable with added risk/uncertainty.

The main variables in the economic analysis are the timing of remedial action, the spatial context and scope of the action, and the key economic parameters such as discount rate and planning horizon (discussed below). Preventive action could be taken now to avoid future damage, or can be postponed allowing existing damages to continue.

The other variable is spatial – the location at which the avoidance or remediation takes place. For each option being considered, at whatever level of interest, the EESA examines the sum of the benefits of the action over the planning horizon, and compares them with the full-life-cycle environmental, social, and financial costs of remediation. As with any decision-making tool, there is uncertainty surrounding key parameters, which are best addressed by sensitivity analysis. EESA should be run under different assumptions to see what the impacts of different parameters on the results are. The goal is to determine the option that provides an environmental, social, and economic optimum for all stakeholders in the majority of likely combinations of parameters. This process is discussed in more detail below.

The EESA can be used to determine an environmentally, socially, and economically sustainable remedial objective, by considering the costs of remedial

alternatives, the benefits of remediation, and the external costs of remediation. Benefits of remediation can be expressed as the value of avoided damage, or prevention of future damages which would have occurred if the remedial action had not been taken [7]. Action which eliminates or reduces damage already incurred is also considered. Each of these elements is discussed in more detail below.

Costs of Remediation

Financial Costs of Remediation The costs of undertaking remediation of contaminated groundwater are well documented and understood, given the decades of activity in this area, particularly in North America and Europe. Remediation costs vary with size of the plume, type of contaminants, and the nature of the geologic and aquifer material and properties. Typically, the later is the intervention, the higher the costs of remediation are likely to be. If the prevention or remedial actions taken produce a secondary impact, it should be included in the analysis as an external cost of remediation (discussed in the following section).

External Costs of Remediation External costs of remediation represent the value of damage done to society and the environment as a result of the remediation process itself. They are “external” in the sense that they are not incurred by the problem holder but by third parties. External costs of remediation can be divided into planned or process-related external costs (which cannot or will not be mitigated against), and unplanned or inadvertent or unforeseen external costs (to which a probability of occurrence can be attached).

Planned external costs may include the landfilling of wastes excavated from contaminated sites, which may result in secondary damage at the new location, and will generate costs to society associated with transporting waste to landfill using heavy goods vehicles by way of increased congestion; impacts on health from air emissions which are produced by remediation systems (such as thermal treatment), noise impacts, and increased probability of accidents [9, 12]. For instance, the external costs of transporting contaminated materials by road in heavy goods vehicles have been estimated at US\$0.78/vehicle-kilometers (v-km) [10]. While there are some more up-to-date estimates

for external costs of transport, this estimate is used here for consistency with the case histories presented here. This can add up quickly. For 10,000 vehicle movements, each of 500-km round trip, using the above unit cost estimate, an external cost of US\$3.9 million is added to the overall cost of remediation. The relevance of this impact can be seen by considering typical private remediation costs for excavation and landfilling of 10,000 t of contaminated soil. A typical remediation program of this size would cost in the order of US\$2 million to US\$5 million, depending on location, contaminant type, tipping fees, and the complexity of the dig. In this example, the expected private or internal cost for remediation using “dig-and-dump” was expected to be approximately US\$3.2 million. Adding US\$0.4 million to reflect the real cost of the remedy represents a 12% overall increase in cost [8]. If clean fill has to be imported to site to fill in the excavation, additional vehicle movements will be required, further boosting the external cost of transport. Furthermore, the other possible external costs of landfilling have not been added, and may be considerable [8].

Examples of other types of planned external costs of remediation are listed in Table 1, below. In general, planned external costs are increasingly being mitigated against. In many jurisdictions, specific regulatory measures are being put in place to ensure that remediation methods which deliberately shift costs from the problem holder to society are reduced or eliminated.

Accounting for unplanned or unforeseen external costs of remediation is of course problematic. Sometimes, despite best planning and care, remediation activities result in creation of a secondary impact to the environment, or to other stakeholders. If the impact is an unplanned or unforeseen result of remediation, for which mitigation measures have not been provided or have not been successful in countering, then the value of this damage is included as an external cost of remediation. Table 2 provides a list of examples of unplanned external costs.

Accounting for unplanned external costs within an economic evaluation of remedial alternatives is not straightforward. For any given remedial approach being considered, the possibility that its implementation may cause additional external damages must be carefully evaluated. In most situations, experienced remediation engineers and specialists should be able

Groundwater Remediation, Environmental and Economic Assessment. Table 1 Examples of planned external costs of remediation

Activity	Secondary effect	Comments
Air stripping of volatile compounds from groundwater, without off-gas treatment	Release of volatile compounds to atmosphere	Still occurs in many jurisdictions, can be mitigated against
Thermal treatment of contaminated soils	Release of CO ₂ and other gases to atmosphere	Greenhouse gas emissions
Permanent geo-sequestering of contaminated groundwater (deep well disposal)	Permanent loss of injected groundwater as a resource	Widely used for difficult and recalcitrant contaminants
Excavation of concentrated source of contamination to protect underlying groundwater results in habitat destruction	Habitat in excavated area destroyed	Mitigation “banking” approaches can be used to offset

Based on Hardisty [8].

to identify possible secondary damages. In all cases, mitigation measures should be put into place to deal with these possibilities. Whatever probability remains of that damage occurring, after the mitigation measures are applied (and the cost of mitigation is added to the overall cost of remediation), should be applied to the value of the damage anticipated should the event occur. Assigning a probability to an eventuality which is being mitigated against is a matter of professional judgment of the remediation team, and should be based on experience, knowledge of the limitations of remedial technologies, and the mitigation measures themselves.

Benefits of Remediation Traditionally, when examining the “economics” of site remediation, the focus has been placed on cost. This has led to a fixation, in many parts of the industry, at many levels, on least cost solutions. However, there has been little consideration

Groundwater Remediation, Environmental and Economic Assessment. Table 2 Examples of unplanned external costs of remediation

Activity	Secondary effect	Example
Remediation causes LNAPL to revert to DNAPL, due to preferential removal of lighter compounds	NAPL sinks, contaminating a new volume of aquifer, worsening dissolved phase problem	SVE (soil vapor extraction) preferentially removes volatile aromatics from an LNAPL containing less-volatile dense compounds
Bioremediation results in creation of daughter products which are more toxic than parent	Toxicity to receptors increases	TCE (trichloroethene) degrades to VC (vinyl chloride), and VC persists in aquifer
Remediation inadvertently increases mobility of contaminant within the aquifer, through alteration of physiochemical properties	Impact on receptors worsens, either due to further spreading of plume, increased mass flux, or more rapid breakthrough	Surfactant flush greatly increases dissolution and mobility of NAPL, which migrated into previously uncontaminated rock
Remediation inadvertently increases mobility of contaminant within the aquifer, through alteration of properties of the aquifer itself	Impact on receptors worsens, either due to further spreading of plume, increased mass flux, or more rapid breakthrough	In situ fracturing of aquifer to enhance NAPL recovery inadvertently allows increased NAPL mobility toward receptors
Remediation compromises adjacent confining layers or geological features	Contaminant is introduced into a hitherto uncontaminated geologic unit	Pumping wells completed across a confining layer, cross-connecting two groundwater-bearing zones

Based on Hardisty [8].

of whether the lowest cost solutions actually yield commensurate benefit for society. Rational, economically balanced decision-making requires that cost be balanced with benefit – the action must be “worth it.” If the lowest cost remedial solution is far greater than the value of achieving the clean-up goal, society is worse off if the remediation is undertaken. Conversely, if there is significant value in achieving a remedial objective, then any method that achieves that clean-up at a cost lower than the benefit which will be realized is a good deal for society. To make such a direct comparison, remedial benefit must be expressed in the same units as the costs of remediation. Without a common unit of measure, decision-makers cannot know definitively if the remediation is actually achieving overall benefit. In the authors’ experience, much of the contaminated site remediation done around the world over the last 20 years has likely not produced an increase in overall human welfare. Conversely, much of the remediation that has been undertaken on major issues has likely been underdone, often because insufficient funds were devoted to the effort in relation to the significant social benefit that could have been realized by the right

level of remediation. Understanding and quantifying the benefits of remediation is the critical step in allowing a more balanced allocation of resources to remediation, and selection of objectives, strategies, and technologies that optimize outcomes.

Private Benefits If the analysis is undertaken at the company (or problem holder) level, where only the costs and benefits that will accrue to the problem holder are considered, then the analysis is a *financial* analysis. When estimating the financial costs and benefits, market prices are used, including the subsidies or taxes that are included in the market price. Financial analysis does not deal with environmental or other social impacts of an investment unless these have a direct implication for the costs and benefits of the problem holder. In essence, financial analysis is what is traditionally done when evaluating remediation. Table 3 presents a selection of benefit categories that can be used in a financial analysis.

External Benefits If the analysis is undertaken for the whole of society, then the analysis used will be an

Groundwater Remediation, Environmental and Economic Assessment. Table 3 Private benefits of remediation

Private benefit	Comment
Increase in property value	Applies to increase in the value of the property owned by the proponent, which results from the clean-up action. Benefit is the net increase in value over the pre-remediation value. In many instances, this can be a major internal driver to remediate contaminated sites on high-value urban properties
Elimination or reduction of corporate liability	Remediation (and possible sale) of a contaminated site may allow an owner to eliminate a financial liability or provision currently affecting its balance sheet. This can be seen as a direct financial benefit to the owner or company. Contaminated site liability provisions are, depending on the jurisdiction, often based either on guesswork, or on an estimate of remediation cost. On that basis, the net position can often be neutral
Public relations value	Remediation of a contaminated site can result in a reduction in ongoing negative public relations, which may result in improved stakeholder relations, lower cost of capital, or perhaps improved financial performance through customer attraction. In practice, it can be difficult to quantify this benefit
Avoidance of prosecution or fines	If remedial action avoids fines or legal action against the company, the costs avoided are a direct financial benefit to the company. However, when a complete economic analysis is done, these costs are not included as benefits, as they are simply transfer payments (the payments are a cost to the company, but a benefit to society (the government), so they cancel each other out)
Health and safety benefits	If remediation of contamination reduces health and safety impacts on workers of the company, then this will be a direct benefit to the firm, in terms of reduced expenditure on ongoing protective measures, increased workforce productivity, lower absenteeism, and lower medical costs borne by the company. Improvements in the workers' own health, and the benefits the workers themselves realize, are counted as external benefits (see below)
Protection of resources used as production inputs	If remedial action protects the quality or quantity of a resource which the company uses in its production, financial benefits may also be realized. For example, if contamination from the facility is making its way into an aquifer that the company uses as a source of production water, and the company has then to treat that water to allow it to be used, remediation may result in the avoidance of some or all of those ongoing treatment costs

Based on Hardisty [8].

economic (or social) analysis. The EESA method focuses on the wider social analysis (which, of course, includes the financial), to examine the costs and benefits that accrue to society as a whole. External benefits of remediation are those that accrue to the rest of society when a problem holder undertakes remediation. If contaminated sites create damage, either because they are not remediated, or because only some of the effects of the contamination are dealt with, then this damage is an external disbenefit (or cost). Table 4 presents a selection of benefit categories that can be used in an economic (social) analysis.

In practice, only some of the external benefits of remediation can be readily quantified and monetized.

The degree to which monetization of benefits is taken depends on the circumstances of the analysis. For more complex, high profile, and serious problems, a greater degree of analysis would be warranted. Wherever benefits can be reliably monetized, they should be. Some of the more important benefit categories listed in Tables 3 and 4, which are particularly important for contaminated sites, are discussed in more detail below.

Remediation of Brownfield Sites: Unlocking Private Benefit As discussed in Tables 3 and 4, one of the most robust ways of examining the economic impact of site contamination is to consider its effects on property value. In many places, the value of property

Groundwater Remediation, Environmental and Economic Assessment. Table 4 External benefits of remediation

Private benefit	Comment
Increase in value of neighboring properties not directly affected by contamination	Increase in the value of neighboring properties which are not physically or directly impacted by the contamination, but are simply affected “by association” with the site, will often results from the clean-up action. Benefit is the net increase in value over the pre-remediation value. The elimination of “blight,” which results from proximity to a site, which in some way diminishes people’s enjoyment of their property (visual disamenity, odor, concern over health impacts), may all depress property values. This effect is widely observed in the literature and in empirical studies of property value [11]. Reduction in blight through hedonic pricing captures a bundle of benefits accruing to those in the affected properties [12]
Increase in value of neighboring properties directly affected by contamination	If nearby properties are physically affected by the contamination – contaminants have been deposited or have migrated from the site to neighboring properties – then remediation of the site itself may also improve the value of the adjacent properties. If the remediation extends to the nearby properties themselves (the company cleans up the neighboring sites also), then the increase in property value of the nearby properties is also an external benefit of remedial action
Health benefits to neighboring residents	If remedial action improves the health of residents living near the site, an external benefit would result. These benefits are real, and can be measured through reduced medical expenditure, improved productivity, or increased income due to improved work attendance. They are not typically captured in a traditional financial analysis. Care must be taken to ensure that a health component is not also included in the property value benefit (double-counting)
Health and recreational benefits to visitors of the area	Remediation of a contaminated site may reduce the real or perceived health impacts on visitors to an area. This can be valued in the same way as the health benefits to residents of the neighboring area. If the site contamination has been affecting the perceived enjoyment of visitors to a nearby recreational area, either by direct contamination, or by “association,” then remedial action can also trigger an increase in wider benefit experienced by the user group
Reduction in ecological damage	In many situations, contaminated land may act as a long-term source of deleterious impact on ecological resources, either within or outside the site. Wetlands, forests, aquatic ecosystems, coastlines, marine habitat, and individual species may be adversely affected by contamination. Elimination of these impacts, and the improvement in the quality and health of the affected ecosystems, can be an important external benefit of remediation, and one which is rarely, if ever, captured in conventional financial analysis
Protection of resources used by others	If remedial action protects the quality of quantity of a resource which is being used (or may be used in the future) by other parties, this would accrue as an external benefit to society. For example, if contamination from the facility is making its way to an aquifer that is used as a source of water for a community, and the community is being affected by that contamination, directly or indirectly, remediation will result in benefits to society. In this example, the benefit could be the elimination of the need for end-of-pipe treatment for the contaminants before distribution of the water to users, or elimination of the need to develop an alternative source of water. For water resources, all elements of the total economic value (TEV) may benefit from remedial action, including nonuse and option values

Based on Hardisty [8].

drives efforts to remediate and then sell on the land. First, there is the value of the site itself. A contaminated property will almost always sell at a discount over the price that could have been fetched if the site were fit for purpose. Groundwater contamination associated with the site can impact on property value in the same way

as soil contamination. In many cases, one of the key benefits of remediation to the owner of a contaminated site is the increase in property value achieved. If the increase in value is greater than the costs of remediation, a profit is realized. This can be a powerful impetus to clean up sites. Different

organizations and individuals have different levels of risk tolerance, and so their willingness to pay for liability reduction can be markedly different. Since market prices are often driven at least partially by perceptions of risk, and these perceptions can vary considerably, regulators have a key role to play in brownfield transactions. Regulatory approval of remedial designs and results can be instrumental in creating comfort in the market that remediation has been successful, thus unlocking site value.

Revenue realized by cleaning up a site accrues to the site owner as an internal or “private” benefit. The benefit of remediation is the difference between the value of the site before remediation and the value after remediation. It can be readily and accurately measured in most countries by market techniques, using property value data.

Blight Reduction: An External Benefit When a contaminated site is remediated, it is not only the site owner who benefits. By removing what might have been a potentially dangerous or hazardous condition, the whole neighborhood benefits. Several recent economic studies have shown that people and businesses experience real economic benefit when a neighboring waste site or polluted site is remediated. This is due to the removal of blight (or disamenity) from the properties in the vicinity of the remediated site. Recent research has found substantial negative effects on property values in areas subjected to transshipment of radioactive wastes in the USA [13]. Another study showed a 35% difference between average selling price of homes located within a 2 mile radius of a low-level radioactive waste site in the USA, to those outside a 2 mile radius [14]. Proximity to landfills, active or closed, has also been shown to have a significant depressive effect on property values. In the USA, decreases in property values between 12% and 25% have been recorded, depending on distance from a hazardous waste landfill [15]. A recent study considered the disamenity costs of landfill in Great Britain [12]. The study considered 11,300 landfill sites and over half a million residential property transactions within 2 miles of those landfills, over the period 1991–2000 (inclusive). Residential property prices were found to be negatively affected within 2 miles of landfills. Across Great Britain, property

values were found to suffer a 7% reduction within 0.25 miles, decreasing with distance to a 1% reduction between 0.5 and 1 miles from the landfill, and 0.7% between 1 and 2 miles from the landfill. In Scotland, however, impacts on property values were greater, decreasing 41% within 0.25 miles, 3% between 0.5 and 1 miles, and 2.67% between 1 and 2 miles away.

Removal of disamenity by remediation will cause average property prices in the affected area to rise. This increase, multiplied by the number of properties affected, can be used as a direct market valuation of disamenity or blight reduction, and as an estimate of the economic benefit which accrues to those stakeholders involved. This benefit will be greatest in dense urban areas with many neighbors and higher property values [8].

Case Histories

Remediating NAPLs in Fractured Aquifers

The problems associated with NAPLs in fractured aquifers have received significant attention in the technical literature over the last decade [16]. Concerns over the impacts of chlorinated solvents on groundwater have led to a significant body of work examining DNAPLs (dense non-aqueous phase liquid) in the subsurface, including in fractured rocks [17]. More recently, the unique behavior and problems associated with LNAPLs (light non-aqueous phase liquids) in fractured aquifers have been studied [2]. The highly heterogeneous nature of fractured systems, combined with the inherent complexity of multiphase flow, typically make characterization and remediation of NAPLs in fractured systems difficult and expensive.

Because this type of remediation can be so challenging and costly, there is a need to understand clearly the justification for spending large amounts of money. Indeed, the remedial methods that are used to remove or immobilize NAPLs in fractured aquifers tend to be complex, intrusive, and energy intensive. This, in turn, brings the sustainability of such efforts into question. Is that level of effort always warranted, given the other opportunities that may be available, and the value of the damage to society, the economy, and the environment caused by the presence of the contamination? This section examines the

application of EESA in setting an optimal remedial strategy for this difficult problem.

Remediation of LNAPL and DNAPL in fractured aquifers is a complex undertaking. DNAPLs may migrate to significant depths, and if spill volumes are large and fracture interconnectivity high, DNAPL may invade progressively smaller aperture fractures with depth. As NAPL fluid pressures increase, matrix invasion may also occur. The vertical migration of LNAPL in fractured aquifers is constrained by the water table, but significant penetration beneath the water table may occur, and lateral migration may occur in directions independent of the hydraulic gradient [2]. Within fractured aquifers, NAPL movement is governed by the geometry of the fracture network (including fracture orientations, densities, interconnectivity, apertures, and wall roughness), capillary pressure and fluid saturation relationships, and the properties of the NAPL (density, interfacial tension, viscosity). Whether dealing with LNAPL, N-NAPL (neutral-buoyancy NAPLs), or DNAPL, significant challenges exist when contemplating remediation. First, characterization of the distribution and behavior of NAPLs in fractured rock is difficult [18]. To develop a deterministic model, fracture networks need to be characterized, major fracture sets identified in the field, and representative fracture parameters determined. The occurrence of NAPL within these fractures then needs to be ascertained, areally and vertically. For DNAPLs, definitive characterization to depth may be problematic [19]. Rarely in practice is a complete characterization feasible [20]. Next, proven techniques for NAPL removal from fractures are few. Pump-and-treat methods, while effective for containment, have proven disappointing for NAPL removal, even when coupled with targeted NAPL recovery pumping and skimming [21]. Recently, more aggressive in situ NAPL removal methods have been field tested, including high vacuum extraction, thermal heating, and surfactant-assisted aquifer remediation [22]. These relatively expensive methods have shown good results in some cases, but have not yet been widely applied in fractured rock environments. Finally, when the understanding of contaminant distribution is sketchy, even the simplest remediation techniques can prove unsuccessful. The combination of new or unproven remedial techniques, incomplete characterization, and complex aquifer and contaminant

distribution conditions makes remediation success uncertain. Within this context, a clear understanding of the financial and broader social, environmental, and economic implications of remediation provides decision makers with the means to select achievable and ultimately valuable remedial objectives [8].

DNAPL in a Fractured Carbonate Aquifer

A variety of chlorinated solvents were disposed of at a site in shallow unlined trenches. The site is situated on a hill, on the outskirts of a small rural town. DNAPL soaked into the surficial layer of fine-grained sediment that covers the site, and in places reached the highly weathered top of the fractured carbonate aquifer below. The groundwater surface at the site is within the bedrock aquifer, which is characterized by low yields and marginal quality from a drinking water perspective. Over 20,000 m³ of NAPL-contaminated soils were excavated from the site and thermally treated, effectively making the site itself suitable for sale. However, groundwater monitoring revealed low concentrations of TCE moving off-site within the fractured bedrock aquifer, toward a downgradient wetland. As with all similar systems, the complexity of the fracture regime at the site makes detailed characterization of the nature and occurrence of groundwater contamination difficult. TCE may or may not be present in selected fractures a few meters apart. Here, EESA is applied at its most basic level, by examining the problem over a longer term perspective, developing quick estimates of the benefits that occur from remediation, and comparing them to the costs of various overall remedial approaches.

Benefits of Remediation A simple analysis identifies the following benefits, which may accrue from remediation at the site. In each case, benefits estimates are overstated to skew the assessment toward active remediation as an outcome. Benefits examined were:

1. Increase in property value at the site itself. The area is rural, and land values in the area are relatively low. Clean, the site is worth about US\$0.50 million.
2. Uplift in the value of surrounding properties through removal of blight. Within a 2 mile radius, the sum of property values is estimated at US\$11 million by area realtors. Applying a 10% blight

Groundwater Remediation, Environmental and Economic Assessment. Table 5 Possible remedial benefits summary

Benefit category	Sum of benefits over 20 years (US\$ million)
Property value	0.5
Reduction in blight in neighborhood	1.1
Aquifer protection	2.5
Wetland protection	0.1
<i>Total possible benefits</i>	4.2

factor, remediation of the site would result in a one-time benefit of US\$1.1 million.

- Prevention of aquifer damage. Using simple modeling, the presence of the dissolved phase plume at the site effectively eliminates about 200,000 m³/year of potential abstraction (even though the aquifer is used only sparsely in the area, and not at all for public supply). At 5% discount rate over 20 years, and assuming a value for water of US \$1.00/m³, this equates to about US \$2.5 m in lost aquifer potential.
- Value of the wetland, to which dissolved phase contaminants may flow. Given the low concentrations expected under worse case conditions, a nominal value of US\$0.1 million is assigned (see the previous example in this entry).

Table 5 provides a summary of the possible benefits of remediation.

Remedial Approach Options Recognizing the difficulties involved in remediating low levels of dissolved phase contamination in the complex fractured rock environment, three main remedial approach options were identified as part of the evaluation process: (1) soil remediation (already completed), with MNA (monitored natural attenuation) for groundwater over 20 years; (2) soil remediation (already completed), with selected in situ treatment of identifiable hotspots in groundwater, and 20 years of MNA; and (3) soil remediation (already completed), with groundwater pump-and-treat to contain and reduce the mass of the dissolved phase plume.

Simple High-Level Environmental and Economic Sustainability Analysis The costs and benefits of each of the three remedial options being examined are provided in Table 6. Although the soil remediation has already been completed, the costs of soil remediation are included in each option to examine how they impact the decision. Nevertheless, none of the options is shown to be economic – they do not result in a net increase in welfare for society. The combination of low property values (small benefits), and a complex and difficult to characterize and remediate aquifer (high costs), produces low BCRs below unity. Even if the value of the aquifer were increased, by raising the value of water from US\$1.00/m³ to US\$2.00/m³ (which is well above current estimates of the TEV of water in this part of the world), the costs of remediation are so high that all of the options remain uneconomic.

Implications In this case, all options are uneconomic. The high cost of remediation of dissolved phase NAPL constituents in fractured rock exceed the benefits to society stemming from remediation. Even the soil remediation alone, which was accomplished with an energy intensive thermal desorption system, was, in retrospect, too expensive for the benefits created. This is a classic example of remediation driven by numerical concentration-based criteria, in which risks have been significantly overstated. A less expensive soil treatment method may have brought overall costs down enough for the proposition to be economic. However, it is clear that further remediation, by deploying either options (2) or (3), is uneconomic and unsustainable. In this case, a technical impracticability (TI) waiver would be both technically and economically justified, leading to selection of remedial approach (1) [8]. In the USA, TI waivers have been issued by the regulatory authorities in such circumstances. This analysis suggests that, in fact, TI waivers are really economic impracticability waivers. With enough effort and expenditure, using existing technology, it is highly probable that the low levels of VOCs at and off-site could have been dealt with eventually. However, this would have required an unreasonable level of expenditure. It is not technology that limits the efforts in most cases, but the willingness to use scarce resources to achieve levels of remediation,

Groundwater Remediation, Environmental and Economic Assessment. Table 6 Indicative costs and benefits, base case (20 years, 5%)

Remediation approach	PV cost soil (US\$ million)	PV cost total (US\$ million)	PV benefit (US\$ million)	Net benefit (US\$ million)	BCR
(1) Soil remediation + MNA	3.0	4.5	2.85	−1.65	0.63
(2) Soil remediation + in situ DNAPL hotspot removal	3.0	9.7	3.6	−6.10	0.37
(3) Soil remediation + pump-and-treat to contain dissolved phase plume	3.0	13.9	4.2	−9.70	0.30

which may not be justified. EESA provides a way of rationally and objectively determining where and to what level remediation is warranted, and where it is not.

Coal Tar Contamination of a Fractured Aquifer

A disused manufactured gas plant facility in the UK is the source of coal tar compounds, including NAPLs, which have contaminated the subsurface. The site lies in the center of a busy town, adjacent a high-value residential neighborhood. The uppermost coarse-grained saturated gravel deposits are extensively contaminated by NAPL, which is very close to neutral density. Groundwater occurs at a depth of about 5 m across the site. A small volume of the NAPL has penetrated a low permeability layer and is present in small aperture fractures within the upper horizon of the underlying carbonate aquifer, to depths of up to 15 m below ground level. The NAPL plume is stable, and occurs mostly as residual saturation. The N-NAPL is composed predominantly of low-solubility PAHs [8].

The aquifer is used for public water supply (PWS), and an operating supply well lies approximately 2 km downgradient of the site. Trace levels of phenolic compounds have been detected at the well over the past several years. Groundwater monitoring and modeling have confirmed that these compounds are migrating from the site in low concentrations. A small culverted river of poor quality runs past the site. The contaminated gravels on-site are in hydraulic connection with the river, but flux of contaminants to the river in the dissolved phase is low, because of the low solubility of the NAPL. The site is derelict, and there is no public access.

Remedial Objective The site owner wishes to remediate the site because of its high value. The regulator and the operator of the PWS well are concerned about contamination of the regional aquifer from contaminants on-site. To date, contaminants detected at the PWS have been at concentrations below drinking water quality standards, and an existing treatment system removes them. However, should the flux of contaminants to the wells increase, there is a risk that the PWS would have to shut down or substantially upgrade the existing treatment system. The objective of the EESA in this example is to: “*select a remedial approach which manages the risks present at the site in the most economic and sustainable way.*” Remediation is to be undertaken sustainably, but what the remediation achieves should also be sustainable in its own right.

Benefits of Remediation The direct financial benefit of remediation to the site owner is the value of the property itself, which can be unlocked if remediation makes the site suitable for hard-covered light commercial use (approximately US\$14.0 million). However, other benefits will also accrue to other parties if remediation proceeds, depending on what the remedial design achieves. Sale and redevelopment of the site will immediately trigger an increase in property values around the site, through removal of blight. As discussed above, this is a well-documented and empirically proven effect. If remedial action significantly reduces the mass flux of contaminants reaching the aquifer, this would prevent future damage to the aquifer and to water production from the PWS (currently pumping 1.0 Mm³/year). Any significant increase in mass flux of contaminants to the well would likely bring it off line. Similarly, removal of NAPL on-site

Groundwater Remediation, Environmental and Economic Assessment. Table 7 Base case benefits summary (5% discount rate)

Benefit category	Sum of benefits over 20 years (2008 US\$ million)
Property value	14.0
Reduction in blight in neighborhood (at 5% blight factor)	3.83
Aquifer protection (base case TEV US\$ 1/m ³)	12.46
River protection (base case TEV US\$ 1/m ³)	3.74
<i>Total Possible Benefits</i>	34.03

would eliminate the source of contamination to the river, whose average annual flow is approximately 0.3 Mm³/year. For the base case analysis, a replacement value of water of US\$1.00/m³ is used. Later in the sensitivity analysis, TEV of water estimates of up to US\$2.00/m³ are used. A summary of possible benefits realized by remediation is provided in Table 7. If all benefit categories could be realized, total benefits would be approximately US\$34 million. However, not all benefits categories will be realized by each remedial approach, as shown below.

Remedial Approach Options Based on the risk assessment discussed above, and understanding the overall social value of preventing damage to the various receptors, and the value of the site itself, a number of remedial approaches were identified to meet the objective. Table 8 lists each option, and describes its basic components and its advantages and disadvantages.

Indicative Remedial Costs Table 9 presents indicative remedial costs for each of the remediation approaches identified above. The remedial approaches are coded to indicate whether they address the risk source (S), or pathway (P) of the problem. Both capital and O&M (operating and maintenance) costs are provided for each approach. External costs of remediation, where applicable, have also been estimated and the basis for those costs are presented (generation of

greenhouse gases from thermal treatment, vehicle use, and power consumption). A base case cost of carbon was set at US\$52/t CO₂e increasing at 2% per year was used, based on the UK government guidance at the time [23]. External costs of landfill were deemed to be included in the capital costs estimates, where applicable, under the assumption that landfill tax effectively accounts for the external social costs of landfill. The costs of road transport disamenity were also estimated assuming that the landfill was 100 km away from the site, and that the heavy goods vehicles used carry an average of 10 m³ of material each. Different planning horizons and discount rates are investigated as part of the sensitivity analysis.

Base Case Analysis Table 10 shows the apportionment of benefits among the various remedial approach options. Apportionment of benefits requires careful judgment of the anticipated effectiveness of the techniques assumed to form the basis for the remedial approach. A zero benefit in the table indicates that the remediation approach does not address the damage to that receptor or does not allow that particular type of redevelopment of the site. The basis for apportionment of benefits must however remain consistent and balanced between options. So, for instance, in Table 10, Option S2 is deemed to eliminate 100% of the damage to the aquifer through aggressive removal of the NAPL in fractured bedrock, using concentrated in situ remediation. In contrast, a system that captures contamination in the surficial groundwater and prevents it from reaching the river (option P1) does not protect the fractured aquifer. P1 does protect the river, however, and therefore accrues that benefit.

Figure 1 shows the base case 20 years, 5% discount rate full social NPVs for each option. The S1 option is superior under base case conditions. Note the contrast with Fig. 2, which shows the same base case NPVs over only 5 years. With a shorter term view, S3 is most economic, because the immediate returns to the problem holder from selling the site now dominate the analysis. Other benefits which accrue to society over time, such as protection of river and aquifer, are deemphasized if only a 5-year horizon is considered. This illustrates the importance of a longer term view, if sustainable outcomes are being sought.

Groundwater Remediation, Environmental and Economic Assessment. Table 8 Remedial approach options

Approach description		Relative cost	Advantages and disadvantages
Source methods			
S1	Partial excavation and removal of main sources. On-site ex-situ treatment. Removal and landfill of remaining material (10,000 m ³). In situ remediation of NAPL in fractured bedrock using chemical oxidation (ISCO) technology	Moderate – High	Pilot trials of ISCO show good success in oxidizing NAPL within fractured rock. On-site treatment focus reduces volume of material for off-site disposal to landfill
S2	Full excavation of contaminants, including into uppermost part of aquifer, use of piling, maximize off-site disposal (40,000 m ³)	High	Quickest and most complete remediation of sources; significant risk of disturbing aquifer during excavation and mobilizing some NAPL into the aquifer, possibly worsening situation at PWS
S3	Partial excavation of NAPL-contaminated materials above water table only (10,000 m ³), leave remaining contaminants in place and monitor	Moderate	Removal of major concentrated sources allows site to be sold with some residual liability; leaves significant mass of contaminant in place
Pathway methods			
P1	Containment in gravels using air-sparge barrier along the downgradient site boundary within gravels.	Moderate	Prevent further off-site migration of contaminants in gravel to River; does not deal with other hazards; perpetual management needed
P2	Hydraulic containment of dissolved phase in bedrock aquifer through installation of a series of capture wells, and operation of a water treatment facility on-site	Moderate	Prevent further off-site migration of contaminants in aquifer to PWS; does not deal with other hazards; perpetual management needed
Receptor methods			
R1	Treat groundwater after abstraction at PWS	Moderate	Directly prevents contaminants from site impacting delivery of water to PWS customers. Treats other contaminants from other sources impacting aquifer also; perpetual management required
Institutional management			
MNA	Monitored natural attenuation. Current data indicate that low levels of dissolved phase contaminants are naturally degrading within the aquifer. No remedial action, monitor groundwater to prove attenuation is occurring.	Low	Does not actively remediate site to any degree; does not eliminate the risk that contaminant flux from the site could increase significantly in future

Based on Hardisty [8].

Remedial approaches MNA and P1 are NPV negative – they do not generate benefits sufficiently large to overcome their costs, even when wider benefits to the whole of society are considered. All of the other remedial approaches generate more benefits overall than

they cost to implement. In this example, cost benefit analysis suggests that excavation and treatment of the bulk of contaminated soil, where the vast majority of the contaminant mass exists, is the most economic single remedial approach evaluated. This is driven by

Groundwater Remediation, Environmental and Economic Assessment. Table 9 Indicative costs for selected remedial approach options

Approach	Capital costs (US\$ million)	O&M costs (US\$ (million)/year)	20 years – 5% PV financial costs (US\$ million)	GHG (tCO ₂ e)	Road transport (v-km)
R1	0.25	0.21	2.87	100/year	–
MNA	0.5	0.10	1.75	10/year	–
P2	3.5	0.35	7.86	200/year	–
P1	3.8	0.40	8.78	100/year	–
P1 – ALT	5.4	0.48	11.38	150/year	–
P1 + P2	6.5	0.55	13.35	300/year	–
S3	9.2	–	7.2	650	200,000
S1	13.9	–	14.4	1,200	480,000
S2	22.0	–	22.0	2,100	800,000

Groundwater Remediation, Environmental and Economic Assessment. Table 10 Base case PV benefits for remedial approaches (US\$ million)

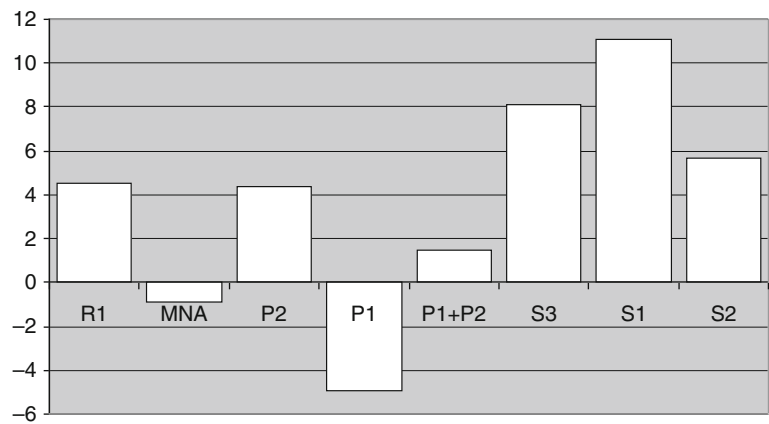
Option	Total financial costs	Benefit – land sale (private)	Benefit – elimination of blight to neighboring properties	Benefit – river protection	Benefit – value of aquifer protection	Total PV benefits
R1	2.87	–	–	–	7.48	7.48
MNA	1.75	–	–	–	0.75	0.75
P2	7.86	–	–	3.74	–	3.74
P1	8.78	–	–	–	12.48	12.48
P1 + P2	13.35	–	–	16.21	12.48	28.69
S3	9.2	7.0 ^a	3.875	1.25	1.25	13.38
S1	13.9	14.0	3.875	3.74	9.35	30.97
S2	22.0	14.0	3.875	3.74	12.48	34.95

^aAssumes that site must be sold at a 50% discount due to the presence of residual deeper subsurface contamination. However, commercial redevelopment can take place and so blight is removed.

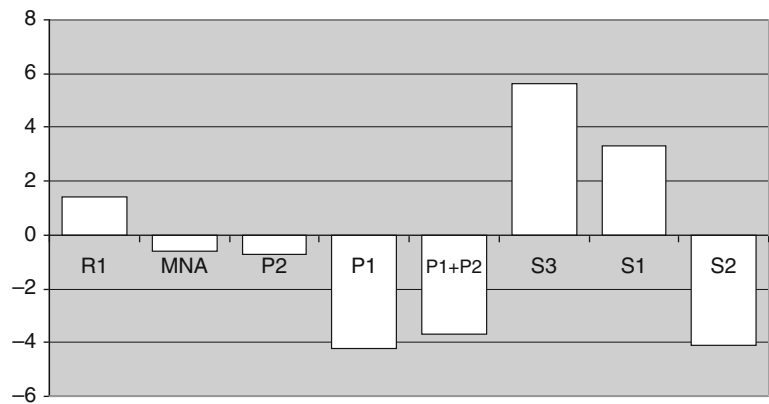
the relatively high value of the property for commercial development, and the fact that all of the neighbors also benefit by elimination of blight on their property values. Source removal will also benefit both the river and the aquifer to varying degrees by removing a continuing source of contaminants.

Under base case assumption, the river's overall value is too small to justify direct action (using the

methods examined) aimed specifically at its protection. In this case, using conservatively high base case valuation parameters for the river, a PV of about US\$3.7 million is estimated. Thus, to be considered sustainable, remediation measures aimed specifically at protecting the river should not cost appreciably more than this figure, in PV terms. If appreciably more is spent on this issue alone, then society as a whole is



Groundwater Remediation, Environmental and Economic Assessment. Figure 1
NPVs of each option under base case conditions, over a 20 year planning horizon (After Hardisty [8])



Groundwater Remediation, Environmental and Economic Assessment. Figure 2
Base case NPVs with a 5 year planning horizon (After Hardisty [8])

worse off. Other projects should be sought which provide higher benefits. If combined with measures intended to manage risks to other, more valuable receptors, river protection can be achieved economically and sustainably. This conclusion is, of course, only valid to the extent that the value of the river assumed for the analysis encompasses all of the services it provides, and that value is stable over time. In this example, it is safe to say that both conditions are being violated. The full extent of the services that natural ecosystems provide to mankind are only slowly being revealed, and the inherent value of these ecosystems will increase over time as population growth, resource extraction, development, and pollution reduce their stock and health. This is

directly reflected in society's willingness to pay (WTP) for those assets. This is explored in more detail below in the sensitivity analysis section to determine if this kind of variability will affect the overall conclusions.

Sensitivity Analysis and Decision-Making Because of the range of assumptions typically required for this type of analysis, sensitivity analysis is an important tool for testing the robustness of conclusions. EESA uses detailed sensitivity analysis to explore the consequences for decision-making overall, and to help guide an optimal decision. For a more detailed discussion of the EESA as a whole and the EESA equiprobable sensitivity

analysis process in particular, the reader is referred to [8]. Benefit values, particularly, should be varied over the widest possible reasonable range, to test how changes in values will affect decision-making. Table 11 shows the ranges of parameters which are varied for the sensitivity analysis in this example.

Figure 3 shows how the NPVs of the various options vary across the full range of values for TEV of water, with all other factors fixed at base case values. This could reflect, for example, a wider anticipated trend toward an increasing value of water, driven by growing demand, resource degradation, and the

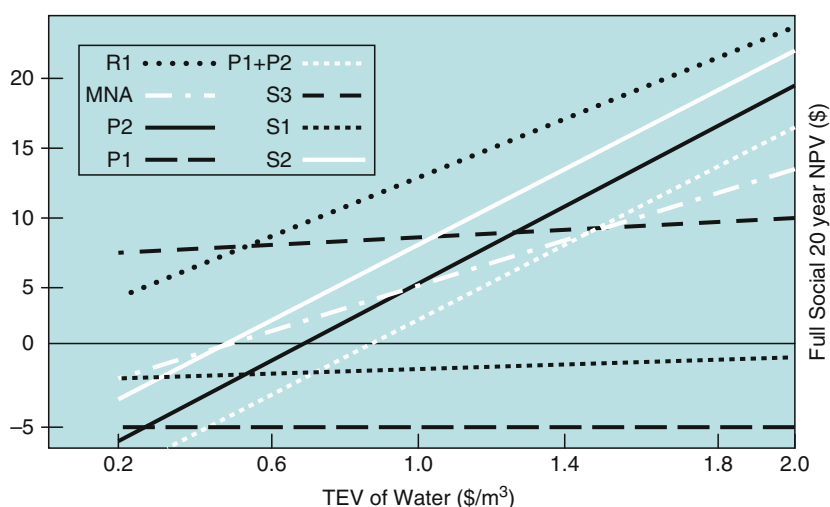
impacts of climate change. Higher values for water have a significant impact on the analysis. More aggressive aquifer remediation efforts such as S1, S3, and P2 become increasingly more economic and sustainable. The more water is worth to society, the more remediation action that protects it can be justified.

Option S1, identified as most economic in the base case analysis, fares well, and is superior under the majority of conditions. This is further illustrated in Fig. 4, which shows that S1 is the best choice under 82.5% of all conditions examined, given equiprobable outcomes. S3 is the best option under 14% of conditions, namely, when the value of water is low, and the revenue from property sale dominates the result.

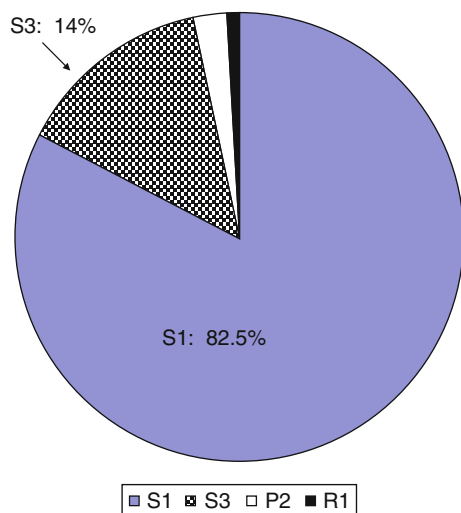
In this example, the site owners could clean up the site at a profit, if no other stakeholder's interests were considered. High property values have driven considerable site remediation over the past decade. However, broader social economic analysis reveals that some of this surplus (profit) could also be used to generate external benefits for the community (and, by extension, the public image of the problem holder). In other words, sustainability in remediation can be identified and justified in an economic context. In this example, several combinations of approaches are in fact possible that would provide wider distribution of the economic benefits of remediation, albeit at higher cost to the problem holder.

Groundwater Remediation, Environmental and Economic Assessment. Table 11 Sensitivity analysis parameter ranges

Parameter	Low value	High value
Property value	US\$5 million	US\$25 million
Blight factor	0%	20%
Total economic value of water	US\$0.50/m ³	US\$2.00/m ³
Discount rate	0.03	0.10
GHG external cost	US\$52/tCO ₂ e	US\$100/tCO ₂ e
External cost of transport	US\$0.60/v-km	US\$2.40/v-km



Groundwater Remediation, Environmental and Economic Assessment. Figure 3 Relationship between net present value and total economic value of water (After Hardisty [8])



Groundwater Remediation, Environmental and Economic Assessment. Figure 4

Chart showing the proportion of instances where specific options were the most economic. S1 was most economic and sustainable over 82.5% of all conditions tested (After Hardisty [8])

Discussion and Conclusions

The case histories presented herein illustrate the application of EESA at very different contaminated sites. While both involve NAPLs in fractured rock aquifer systems, the results from a decision-making perspective are quite different. At the MGP site, high urban property values combined with the risk of impacts on a major aquifer meant that remediation of NAPL in fractured rock, although very expensive, is economic – it yields overall benefits to society as a whole. In contrast, the rural chlorinated solvent problem involves a poor yield, little-used aquifer, and low concentrations of dissolved phase contamination. Remediation of the low levels of TCE in the fractured aquifer is not economic, and not sustainable. The already-completed soil remediation has removed the continuing source of future dissolved phase contamination, and renders the site suitable for limited commercial use. However, further remediation cannot be justified, based on the benefit to be gained. In both cases, active remediation of NAPL in complex and highly heterogeneous fractured rock is very expensive (over US\$4 million and almost US\$10 million, respectively). However, the real

economics of the two cases are very different, and the optimal decision over a wide range of future possible values for key external assets, reflects this.

In general, active groundwater remediation may be beneficial and warranted if property values are high and are expected to rise, if aquifer use value is high, if alternative water sources are scarce and if important nonuse values are threatened (such as groundwater recharging a sensitive and valued wetland ecosystem), and if ecological assets at risk are unique and valued by society. In many cases, the anticipated benefits of remediation in complex technically challenging environments (such as fractured aquifers) will not justify the level of expenditure required to realize those benefits, with currently available technology.

Future Directions

This approach requires valuation of economic externalities. To allow benefits to be more accurately quantified, studies are required on valuing aquifers, both in the USA and Europe. Supporting this, more quantitative work on establishing estimates of the TEV of water, in various conditions (based on rainfall, population density, demand, and so on) would be very useful. Analysis of the type presented herein also suggests that continued research is needed to bring down the financial costs of groundwater remediation, particularly in more complex subsurface environments.

For too long, decision-making about remediation has been carried out with only half of the required information. Legislation and regulatory standards have driven problem holders to seek the cheapest possible way to meet the applied standards. The real overall benefits of groundwater remediation, however, have been largely ignored. EESA provides a robust way to examine the benefits to all parties from remediation, and compare the overall economic sustainability of a variety of options across a range of possible future conditions to determine an optimal decision that will stand the test of time. Application of this type of approach, however, is in its infancy. Development of improved policy, which mandates a wider perspective on remedial decision-making, would help to avoid the extremes in remediation which destroy value in society: spending too much on a problem that does not warrant it and spending too little when valuable groundwater resources are being seriously damaged.

Bibliography

Primary Literature

- Hardisty PE, Bracken RA, Knight M (1998) The economics of contaminated site remediation: decision making and technology selection. *Geol Soc Eng Geol Sp Pub* 14:63–71
- Hardisty PE, Wheeler HS, Johnston PM, Bracken RA (1998) Behaviour of light immiscible liquid contaminants in fractured aquifers. *Geotechnique* 48(6):747–760
- Goist TO, Richardson TC (2003) Optimizing cost reductions while achieving long term remediation effectiveness at active and closed wood preserving facilities. In: *Proceedings, NGWA conference on remediation: site closure and the total cost of cleanup*, New Orleans, Nov 2003
- Perakami MP, Donovan PB (2003) A remedial-goal driven, media-specific method for conducting a cost benefit analysis of multiple remedial approaches. In: *Proceedings, NGWA conference on remediation: site closure and the total cost of cleanup*, New Orleans
- Hardisty PE, Kramer E, Ozdemiroglu E, Brown A (1999) Economic analysis of remedial objective alternatives for an MtBE contaminated aquifer. In: *Proceedings of hydrocarbons and organic chemicals in groundwater conference, NGWA, Houston*
- Pearce DW, Warford JJ (1993) *World without end*. Edward Elgar, London
- Hardisty PE, Ozdemiroglu E (2005) *The economics of groundwater remediation and protection*. CRC Press, New York
- Hardisty PE (2010) *Environmental and economic sustainability*. CRC Press, New York
- Hardisty PE, Arch S, Ozdemiroglu E (2008) Sustainable remediation: including the external costs of remediation. *Land Contam Reclam* 16:4
- Maddison D, Pearce DW, Johansson O, Calthorp E, Litman T, Verhoef E (1996) *The true costs of road transport*. Earthscan, London
- Hardisty PE, Ozdemiroglu E (1999) Costs and benefits associated with remediation of contaminated groundwater: a review of the issues. *UK Environment Agency Technical Report P278*
- DEFRA (2003) *A study to estimate the disamenity costs of landfill in Great Britain*. Final report by Cambridge Econometrics with EFTEC and WRc
- Gawande K, Jenkins-Smith H (2001) Nuclear waste transport and residential property values: estimating effects of perceived risks. *J Environ Econ Manage* 42:207–233
- Payne B, Olshansky S, Segel T (1987) The effects on property values of proximity to a site contaminated with radioactive waste. *Nat Resour J* 27:579–590
- Hirshfeld S, Vesilind PA, Pas EI (1992) Assessing the true cost of landfills. *Water Manage Res* 10:471–484
- Mackay DM, Cherry JA (1989) Groundwater contamination: pump-and-treat remediation. *Environ Sci Technol* 23:6
- Cohen RM, Mercer JW (1993) In: Smoley CK (ed) *DNAPL site investigation*. CRC Press, Boca Raton
- CL:AIRE (2002) *Introduction to an integrated approach to the investigation of fractured rock aquifers contaminated with non-aqueous phase liquids*. Technical Bulletin TB1
- Guswa JH, Benjamin AE, Bridge JR, Schewing LE, Tallon CD, Wills JH, Yates CC, LaPoint EK (2001) Use of FLUTE systems for characterisation of groundwater contamination in fractured bedrock. In: Kueper BH, Novakowski KS, Reynolds DA (eds) *Conference proceedings of fractured rock 2001*, Toronto
- Lane JW Jr, Buursink ML, Haeni FP, Versteeg RJ (2000) Evaluation of ground penetrating radar to detect free-phase hydrocarbon in fractured rocks – results of numerical modelling and physical experiments. *Ground Water* 38:6
- Schmelling SG, Ross RR (1989) Contaminant transport in fractured media: models for decision makers. *EPA Superfund Issue Paper EPA/540/4-89/004*
- Taylor TP, Pennell KD, Abriola LM, Dane JH (2001) Surfactant enhanced recovery of tetrachloroethylene from a porous medium containing low permeability lenses: 1. experimental studies. *J Contam Hydrol* 48:325–350
- DEFRA (2007) *The social cost of carbon and the shadow price. What they are and how to use them in economic appraisal in the UK*. UK Department of Environment, Food and Rural Affairs, London

Books and Reviews

- Boardman AE, Greenberg DH, Vining AR, Weimer DL (2010) *Cost-benefit analysis: concepts and practice*. Pearson, Prentice Hall, Upper Saddle River
- Brouwer R, Pearce D (eds) (2005) *Cost-benefit analysis and water resources management*. Edward Elgar, Cheltenham
- Pearce DW, Atkinson G, Mourato S (2006) *Cost benefit analysis and the environment – recent developments*. OECD, Paris
- Tientenberg T (2003) *Environmental and natural resource economics*. Addison Wesley, London

Groundwater Salinity Due to Urban Growth

J. J. CARRILLO-RIVERA¹, S. OUYSE²

¹Instituto de Geografía, UNAM, CU México, DF, Mexico

²Cadi Ayyad University, Marrakech, Morocco

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Groundwater Movement in Flow Systems
 Importance of Groundwater for Urban Development
 Groundwater Quality Response
 Geological Framework
 Climate Framework
 Water Quality Response to Groundwater Extraction
 Vertical Flow Control to Prevent Water Quality
 Deterioration
 Future Directions
 Bibliography

Glossary

Aquifer unit Is a geological formation, part of a formation, or a number of formations that provide water substantially and in adequate quality for the expected usage.

Basin Is often referred to as the drainage basin or the surface area where rainfall is gathered with a common discharge outlet; it is considered to have no additional inflow or outflow.

Flow systems Are manifested by the presence of flows with contrasting hierarchy (local, intermediate, and regional) whose components may be clearly defined from field evidence in conjunction with modeling of groundwater hydraulics, geochemistry, geomorphology, edaphology, as well as on natural vegetation response.

Groundwater vulnerability to contamination Is the tendency or likelihood for a contaminant to reach a specified position in the groundwater system after introduction at some location above the uppermost aquifer.

Hydraulic conductivity Is the rate of water that a unit volume of aquifer material may allow through under a unit hydraulic gradient; such value is a function of its degree of saturation, attaining its maximum at 100% saturation, and is also a function of water density.

Local flow system Is one that is defined traveling a short distance and depth, recharging and discharging in the same valley.

Regional flow System Is generated (recharged) at the highest elevation and then travels along till basement depth and discharges at the lowest position in the region.

Urban sprawl Is a multifaceted concept, which includes the spreading of dwelling construction

outward of a city and its suburbs to its outskirts to low-density and becoming an auto-dependant development on rural land, high segregation of uses (e.g., stores and residential), and various design features that encourage car dependency.

Vertical flow Is groundwater traveling downward (as in a recharge area) or upward (as in a discharge) area. This component of flow may be also identified when extraction is carried out at shallow levels and groundwater ascends, induced by extraction.

Water balance As applied to a watershed (basin), is referred to as the application of a lumped parameter approach that lacks consideration of areal variations of precipitation, evaporation, transpiration, recharge, runoff, and change in storage as well as the processes involved.

Water quality change Is the chemical evolution in obtained water quality in borehole with extraction time.

Wellhead protection area Ideally is the entire groundwater inflow area for the well; often, this inflow area is too large to be managed effectively, so a smaller area near around the well is often chosen and delineated by flow lines basically incorporating horizontal flow.

Definition of the Subject and Its Importance

The population of the world is increasing at different rates according to the period of time and geographical position; however, at present, the number of cities that have a population larger than one million are estimated as more than 336, while in the 1950s, 83 cities were in that range. One such urban development has been Mexico City, which has more than 18.6 million population, which could be considered even a larger number (nearly 28 million) if the adjoining urban sprawl is included. Cities have often solved their water supply requirements by means of groundwater extraction. Usually, in ancient times, cities were founded on sites where water was directly available. Water supply works tapped springs and surface water. Later, when the number of inhabitants and economical activities increased, shallow dug wells were the solution to cope with water needs. As historical time went by, the population and economical activities of cities have definitely increased; however, at the turn of the 1900s, many cities experienced a rapid

growth that has evolved to achieve their present size. The consequent need for additional water sources has been met with the assistance of groundwater.

Initially, at the second half of the 1800s, groundwater extraction in Mexico City was made on the surrounding plain by means of boreholes that provided artesian yields with good quality water in sufficient quantities to comply with existing local requirements. However, the limited supply to cope with the continuous increase of water needs led to additional extraction by boreholes. Later, by the end of the first half of the twentieth century, many boreholes were no longer free-flowing artesian waters, so the acquisition of the turbine pump started a new extraction period. At that time, the prevailing view was that water quality would basically remain constant with extraction time due to calculations that included a hydrogeological conceptual model known as the “the water balance,” applied in Mexico from the middle of the 1960s. This conceptual model included an important assumption based on the presence of purely radial horizontal flow to the extraction borehole as defined by Theis [23], a condition that was supposed to prevail in existing extraction boreholes.

Initially, boreholes were skunked to shallow depths in the vicinity of 100 m, just touching the top of the full aquifer thickness. At present, deeper boreholes may reach 400 m but are still far from tapping the full thickness of the aquifer, which is estimated to be of more than 3,000 m.

Originally, the deterioration of obtained groundwater quality was solely expected to occur from spillage on the soil surface of foreign substances (i.e., gasoline leaks, leachates from rubbish pits, improper management of sewage water, etc.), not from substances naturally existing in the groundwater. The common prevailing wellhead protection area was considered to be the only source responsible for any negative change in the obtained quality of the groundwater supply. Often, any lack of effect of lateral flow was overcome by the presence of vertical ascending flow from extraction boreholes.

Vertical components of groundwater flow have been overlooked as well as their meaning in producing related environmental impacts [7]. Vertical ascending components of flow become paramount when viewed in the context that usually the screen of the extraction borehole faces less than 1/10 of the full aquifer thickness. Below this extraction depth, groundwater flow

systems with different water quality travel below the local flows that were initially captured by the boreholes bottomed at shallow depth. Regional flows with both high temperature and salinity load travel at depth and are represented by a thermal spring reported since pre-Spaniard times: Peñón de los Baños (44°C; Cl, 650 mg/L; Li, 1,550 µg/L). Groundwater belonging to regional flows is being slowly but steadily induced to extraction levels, changing the initial water quality. Additionally, there is an increase in the disappearance of springs due to extracting water from boreholes at the spring site. The recognition of the presence of such flow systems is very important. As population and economic activities increase, so will the rate of extraction, which in turn will change the obtained water quality over time.

Introduction

Groundwater may provide feasible alternative solutions to water needs for many growing cities' development worldwide. The definition of groundwater functioning is essential when looking for water accessibility in terms of quantity and quality in both time and space. Such knowledge could assist in providing an adequate definition of water sources to reduce the risk of salinity enhancement of extracted groundwater due to urban growth. Establishment of adequate groundwater management policies needs to be in accordance with the groundwater dynamic in the environment and an understanding of existing flow systems.

The understanding of groundwater functioning is most important in regions where there is an increase in water needs and where shortages in the water supply often affect available infrastructure. The limiting conditions for appropriate development of big cities in many regions could be solved under the discipline of land planning. Proper space planning refers to the methods used by the public sector to influence future distribution of people and economic activities in spaces at various scales. Land-planning processes that are undertaken at local, provincial, and regional levels require covering a broad range of issues, such as housing, commercial and other nonresidential facilities, roads and transportation, schools and utilities, as well as farming. What is meant to be considered proper space planning implies that the proposed land use is in agreement with what the

land “attitude” and its environment could provide without putting in danger its sustainability.

Ideally, socioeconomical and environmental development program should be assisted by an adequate political decision-making to include the definition of the groundwater flow systems. Required actions to increase development involve a responsible and thorough knowledge of the nature of the territory, not only surface elements but also groundwater availability and its underground links with neighboring surface drainage basins. The traditional water availability analysis carried out for additional development in Mexico is related to the “*water balance*” which usually involves hydrological factors that are difficult to obtain through the application of the scientific method. This is due to (1) simplification of natural processes involved; (2) use of estimates of variables not directly measured, such as runoff, precipitation, evapotranspiration, groundwater extraction, infiltration, and groundwater recharge; and (3) presumption of immediate infiltration of local precipitation that is reflected in an ascending water table. Additionally, prevailing geological conditions of Mexico, and in particular those related to the location of Mexico City, imply the presence of aquifer units with thicknesses beyond 3,000 m and with horizontal extent of about 100,000 km². Therefore, most surface drainage basins usually have an underground hydraulic connection with near or faraway surface drainage basins [24]. A starting point with most computed water balance carried out on a surface drainage basin generally fails to represent a consistent scenario with groundwater functioning, and as a result, any groundwater extraction program frequently would fail to consider the expected water quality to be obtained by boreholes over time.

From the perspective of safeguarding obtained water quality, the concept of “borehole protection area” (see <http://www.state.nj.us/dep/njgs/whpaguide.pdf>) in complex hydrogeological systems lacks the required guidance of particular needed measures of protection to preserve obtained water quality and to recommend appropriate monitoring programs for borehole production. Under purely horizontal flow conditions, the sources of water quality variations are solely expected to arrive from the surface and not from beneath the tapped aquifer unit. The concept of groundwater vulnerability to contamination has limited value in protecting existing groundwater supply if

every alteration of water quality is solely expected to occur from above ground level, not from beneath the aquifer unit where the water is being tapped.

A different approach is therefore needed to define changes in the obtained groundwater quality over the long term. A better understanding of groundwater functioning will facilitate the process. Additionally, the inclusion of socioeconomic and political variables has an equal importance in preventing effects resulting in extracted water quality deterioration if extraction and demand activities are properly planned in agreement with the groundwater flow regime. A constant increase in extraction yield requires an interdisciplinary understanding to assist decision-makers in reaching a better assessment on how concepts such as a specific land-use planning may change water quality and availability. A particular policy of land planning on a territory with certain water availability could develop a sustainable use according to the knowledge of water sources’ functioning and their natural limits. Conversely, without a clear perception of the functioning of water sources, decision-makers are very limited in the understanding of impacts of their decisions regarding specific growth of related cities’ needs. In Ref. [7], associated environmental impacts due to a lack of attention to the functioning of groundwater flow systems in México are identified.

In both semiarid and temperate regions, groundwater sources have become crucial to maintain social and economic development [12, 14, 17]. Fresh groundwater is available in many areas as recently infiltrated rainfall manifested in local groundwater flow systems; therefore, it is common sense to improve strategies for its adequate management since it usually remains the lower-cost option for drinking water. This low-cost option is favored not only financially but also environmentally, especially when water transfers among neighboring basins or surface water sources have become polluted. Another alternative often used is to treat the extracted groundwater. However, such a procedure has proved to be inadequate in the case of the control of the continuous inducement of saline waters of Mexico City which exceed local treatment methods and capacity designed to cope with strict levels of salinity. One important issue related to this inefficient rate is linked to water quality changes with extraction time. The quality of the inflow water to the plant will hardly

meet design specifications, and consequently, treatment plants may fail to meet drinking water production standards.

Groundwater Movement in Flow Systems

Groundwater moves in 3-D from its recharge to its discharge area (i.e., river, wetland, lake, spring, sea) in flow systems traveling through paths of different lengths and depths resulting in local- to regional-scale flow systems [25]. The presence of flow systems with contrasting hierarchy may be clearly defined from field evidence in conjunction with modeling of groundwater hydraulics, geochemistry, geomorphology, edaphology, as well as natural vegetation response. Contrasting geomorphological characteristics give groundwater a distinctive flow system with particular physical functioning [25], resulting in specific flow directions: vertical downward in recharge areas, horizontal in transit areas, and vertical upward in discharge areas (Fig. 1). Groundwater flow systems are constantly linked; therefore, it is paramount to understand their functioning to manage them adequately in agreement with their environmental conditions to reduce the impacts of human water usage and land-use changes.

Regions have distinct chemical and physical characters for different flow systems, especially where thick and extensive porous and permeable geological units are present above a basement rock. Usually, different temperature (cold in recharge areas, higher in the discharge area) and chemical conditions (i.e., water with high dissolved oxygen, low pH, and low salinity in recharge areas; and low dissolved oxygen, high pH, and high salinity in discharge areas) are characteristic of existing flow systems. Groundwater flow systems travel in the prevailing geological framework structure along the horizontal and vertical directions including the composition of the geological environment they traveled across. Local flows travel short distances at small depth, while those that initiate their traveling path in a basin and discharge in a neighboring basin containing at least one local flow constitute an intermediate flow (Fig. 1), all of which may be contained by a regional flow. A sketch of the location of Mexico City as related to the prevailing flow systems is represented in Fig. 1 as the current depth of extraction boreholes.

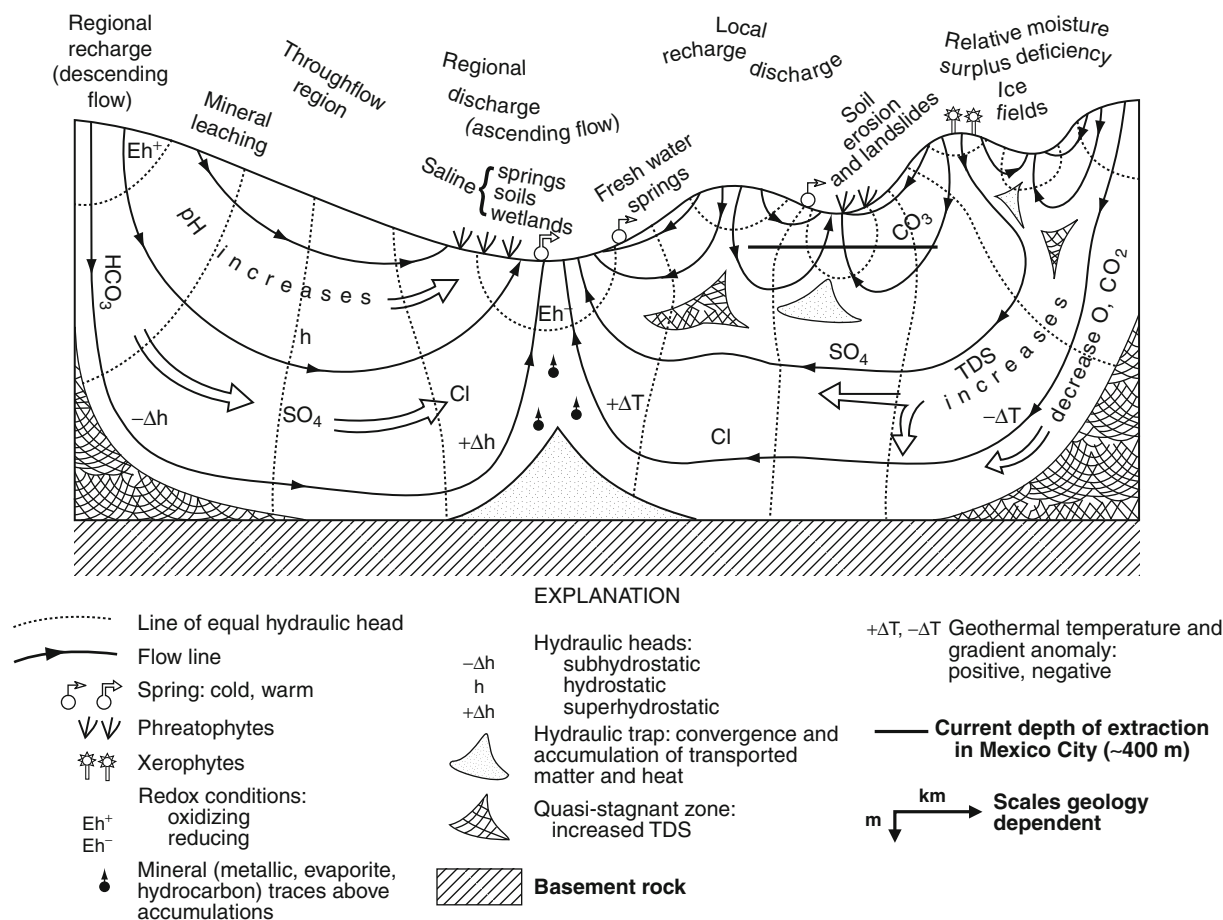
In Mexico, as a result of the thickness and extent of geological formations that constitute the aquifer unit as well as the variability of climatic characteristics, regional flows travel several hundreds of kilometers, often flowing from a temperate climate region to one with arid to semiarid characteristics, where urban growth and related economical development rely on flows of this nature as the city of San Luis Potosí [6]. Such development has been in progress without proper consideration being given to water availability not only in quantity terms but also regarding its quality.

Importance of Groundwater for Urban Development

The Mexico City Case

A large groundwater storage capacity becomes important when urban development exceeds surface water availability. However, the awareness of groundwater functioning is necessary to be understood and to incorporate it in any development plan of megacities as Mexico City (and other major cities worldwide), which is characterized by thick and extensive aquifer units, to avoid inducing inflows of groundwater systems that might deteriorate with time the quality of its originally obtained groundwater supply.

From its establishment, water needed for the development of Mexico City was originally supported from groundwater sources (springs). As time advanced, both development and population increased, and resulted water supply needs were mainly overcome with groundwater. Extraction had a rather stable increase from 1900 to the 1940s (from ≈ 2 to ≈ 4.5 m³/s, respectively); however, from the 1940s, a rampant increase in water needs reached a claimed water supply in excess of 60 m³/s by the end of the twentieth century, where some 76% was obtained from groundwater sources (Mexico basin $\approx 67.0\%$ and Lerma basin $\approx 9.4\%$) and 23.6% from surface sources (Mexico basin $\approx 2.3\%$ and Cutzamala basin $\approx 21.3\%$). This extraction history has produced a drop in water table level from artesian conditions to a depth of more than 50 m; a consequent response of the system to such increase in groundwater extraction is believed to be a key factor to the produced water quality change.



Groundwater Salinity Due to Urban Growth. Figure 1

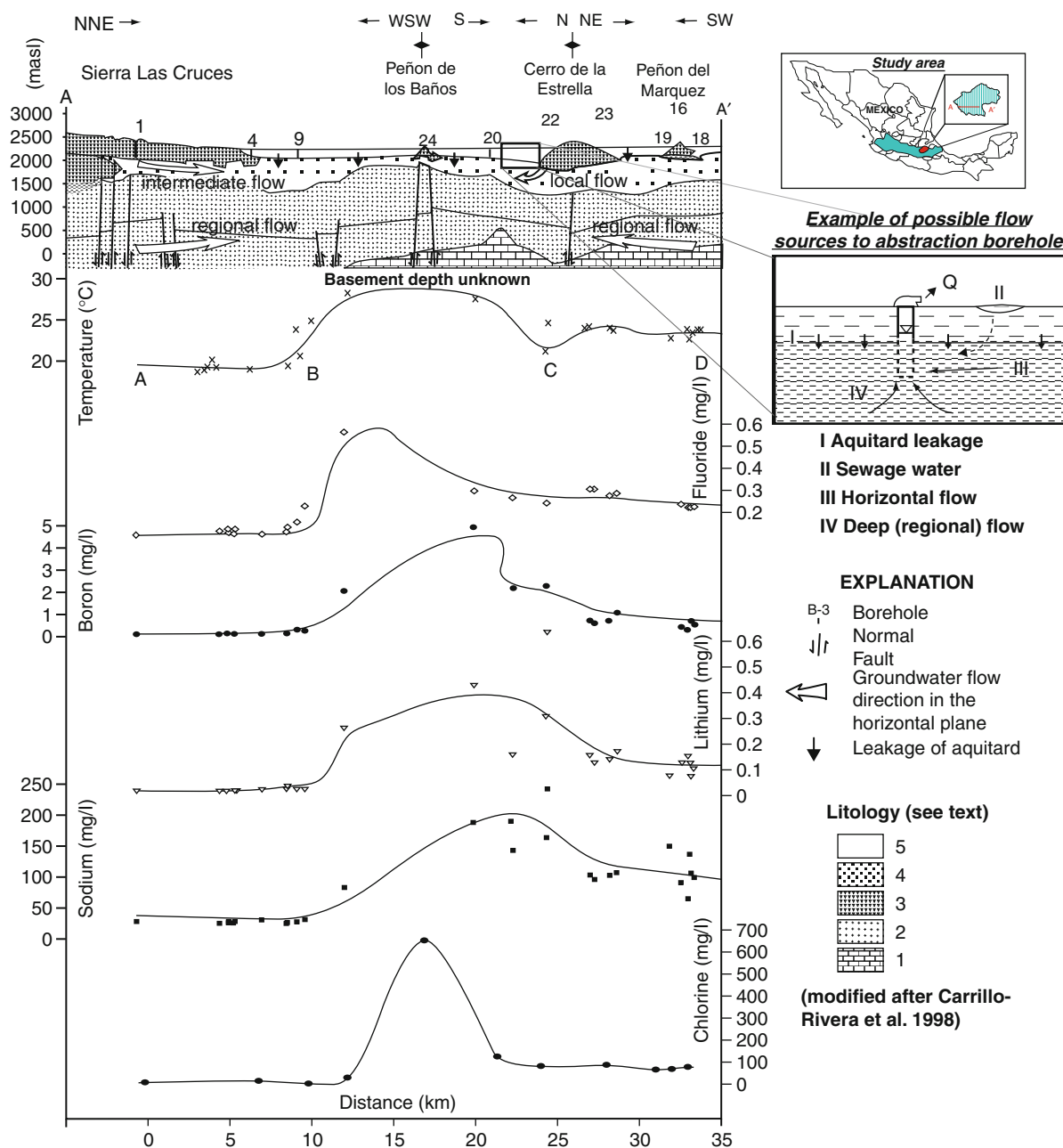
Gravity groundwater flow systems developed in an unconfined drainage basin defining local, intermediate, and regional flow systems (Adapted from [26]). The location of the groundwater extraction for Mexico City is indicated by the depth of boreholes as related to the full aquifer unit

Groundwater Quality Response

Mexico City is located on a plain within a natural closed basin where its water supply since the beginning of the twenty-first century is about 80% supplied by groundwater. The groundwater supply has been of good quality generally for almost a century, from the late 1840s when the first borehole was constructed to the early 1970s when the extracted water quality drastically changed. Initially, groundwater extraction by artificial means was obtained mainly through artesian boreholes at depths of about 100 m. As technology advanced, construction of deeper boreholes was feasible, and the turbine pump was employed to solve the

nonoverflowing conditions; deep boreholes went down to about 400 m, which is their current present depth beneath the foundations of the city. Reported salinity problems from groundwater in the northwest of Mexico City started by 1976, with a reported increase of about 300% in historic salinity; for example, Cl, Fe-Mn, and TDS increased from 32 mg/L, 0.1 mg/L, and 230 mg/L, respectively, to 55 mg/L, 0.45 mg/L, and 750 mg/L, respectively.

An understanding of the response of the chemical groundwater quality under the continuous increase of extraction in the Mexico basin (Fig. 2) may be reached when considering that geological framework and



Groundwater Salinity Due to Urban Growth. Figure 2

Groundwater quality evolution along a section located in the south of the Mexico basin. Inset: Location of the basin in the Trans-Mexican Volcanic Belt

current climatic conditions are directly responsible for the presence of contrasting groundwater flows traveling at different depths with different path lengths. Increased extraction with partially penetrated boreholes also contributes.

Geological Framework

Geologically, Mexico City is located on a plain in a closed basin located on a graben structure developed during the Oligocene, where a thick succession of

volcanic and lacustrine materials was deposited. During the Oligocene, the basin drained to the south. This drainage outlet was closed during the Pleistocene from a series of volcanic activities [11] that formed the Chichinautzin Sierra to the south. The closure was part of an intense regional volcanic activity that formed the Trans-Mexican Volcanic Belt, which now covers Central Mexico from east to west with a length of about 1,100 km and some 400 km in width. The Mexico basin is located at its center (Fig. 2). The extrusive events occurred from 700,000 years before present till the Quaternary; during the later stage, volcanism allowed a substantial heterogeneous layer of ash, interbedded with extensive alluvial and lacustrine deposits, to accumulate in the various lakes formed after the closure of the basin. The full volcanic succession consists of a Middle Tertiary volcanic unit that includes clastic material as well as basalt, rhyolite, and andesite rock units with a collective thickness in excess of 3,000 m. These units are partially covered by Lower Pliocene deposits characterized by lacustrine and pyroclastic material with a total thickness of 600 m. These rocks are overlaid by Plio-Quaternary rocks consisting mainly of basaltic andesite, andesite, and pyroclastic material that dominate the highlands and are interstratified with contemporaneous alluvial and lacustrine deposits; their total thickness on the plain is in excess of 1,000 m. Recent volcanic episodes and regional tectonic events have produced an intensive vertical fault structure identified to be crossing beneath the plain [11].

These volcanic rocks overlie a regional Cretaceous sequence of calcareous deposits that consist of limestone, sandstone, and shale outcropping beyond the southern and northern basin limits; this deepest and oldest identified geological unit with a total known thickness is at the center of the basin in excess of 1,500 m. Deep drilling encountered the top of this unit at 1,581 m below ground surface. Karstic and fractured features are reported based on partial and total loss of drilling fluid and the drop of about 8 m of the drilling tools [28].

Recent deposits cover the entire floor of the plain, comprising some 600 m of fluvial and alluvial deposits. A lens of fine material (microfossils, volcanic ash, and, to a minor extent, lacustrine clay interbedded with sand, silt, and occasionally gravel) is included within these

deposits; it outcrops in Lake Texcoco, underlying the rest of the plain except at the margins; in general, these deposits contain [19] high Na (up to 9,200 ppm) and Mn (up to 1,920 ppm). Total thickness increases gradually from the edge of the plain to Lake Texcoco, where it attains approximately 300 m. This unit forms a main aquitard saturated with water of heterogeneous quality, which locally acts as confining or semiconfining unit to the aquifer material beneath. This aquifer unit includes the extrusive and sedimentary rocks that extend regionally over Central Mexico with a thickness of more than 4,000 m, and it rests above a basement rock that has not been physically described.

The Mexico basin is one of the largest of a series of closed basins located in the Mexico Trans-volcanic Belt [28]. The basin has a drainage area of about 9,600 km²; however, the actual area covered by the city proper and its sprawl area incorporates an extent of about 2,400 km². The abrupt relief of the surrounding mountains, with altitudes in excess of 5,000 m amsl, slopes toward the flat-lying center of the basin to an altitude of approximately 2,230 m amsl.

Climate Framework

The present climatic conditions where Mexico City is located are those of subtropical but with a significant variation in the mean annual precipitation across the basin. Precipitation ranges from 1,800 mm in the southern Sierra Chichinautzin, 1,100 mm in the west (Sierra Las Cruces), and 1,200 mm in Sierra Nevada, to 600 mm toward the remnant Lake Texcoco toward the center of the basin. The rainy season is mainly in the summer months. The mean annual temperature is from 12° to 14 °C in the sierras of Chichinautzin and Las Cruces and 15 °C in Lake Texcoco site. The potential annual evapotranspiration is around 1,400 mm. Rainfall produces negligible surface runoff in the permeable rocks of the Sierra Chichinautzin to the south of the city. In contrast, substantial runoff is generated in Sierra La Cruces to the west of the city, a condition that has caused severe historical flooding.

Water Quality Response to Groundwater Extraction

Mexico City reportedly extracts more than 50 m³/s of groundwater from its basin to supply differing water

needs. A plan for actions for positive environmental response resulting from groundwater extraction should rely on an understanding of the flow systems to assist in planning and control of groundwater extraction by defining which particular flow is more vulnerable to intensive local water extraction rates in time and space.

Four real cases of the Mexico City basin [8] are examined; the interference of particular nearby boreholes is not incorporated. The first case presented by borehole 3 in Fig. 3 shows pumping (and recovery test) data that illustrates the local confining characteristics of the geological environment manifested as the classical drawdown–time response; the nature of the horizontal flow is relatively evident by stable values of TDS and temperature of obtained water with extraction time.

The case of borehole 8 shows a response traditionally considered to be the only possibility of chemical water quality change in this basin, portraying the acknowledged leaky effect from the semiconfining unit located above the level of groundwater extraction. There is an inflow of cold water, and withdrawal from this unit produces water with lower temperature. As extraction progresses, inflow water is expected to reach observed previous values (pv) when the borehole was stopped to initiate the test, at about 21°C. Reduction of TDS shows a change in the chemical characteristics of inflow water due to leakage from the semiconfining unit.

Drawdown–time data in borehole 23 would appear to also suggest a standard leaky response where the drawdown rate is diminished along the duration of the test; however, the increase of groundwater temperature may suggest that water is induced from intermediate or regional flows traveling beneath.

The response of the increase in TDS and constant temperature of borehole 12 suggests that obtained water is derived from standard (semi)confined conditions, where aquitard material above and below the tapped aquifer strata provides a steady response in terms of salinity values. Water is derived from the same shallow depth, keeping temperature at a steady value of 21°C.

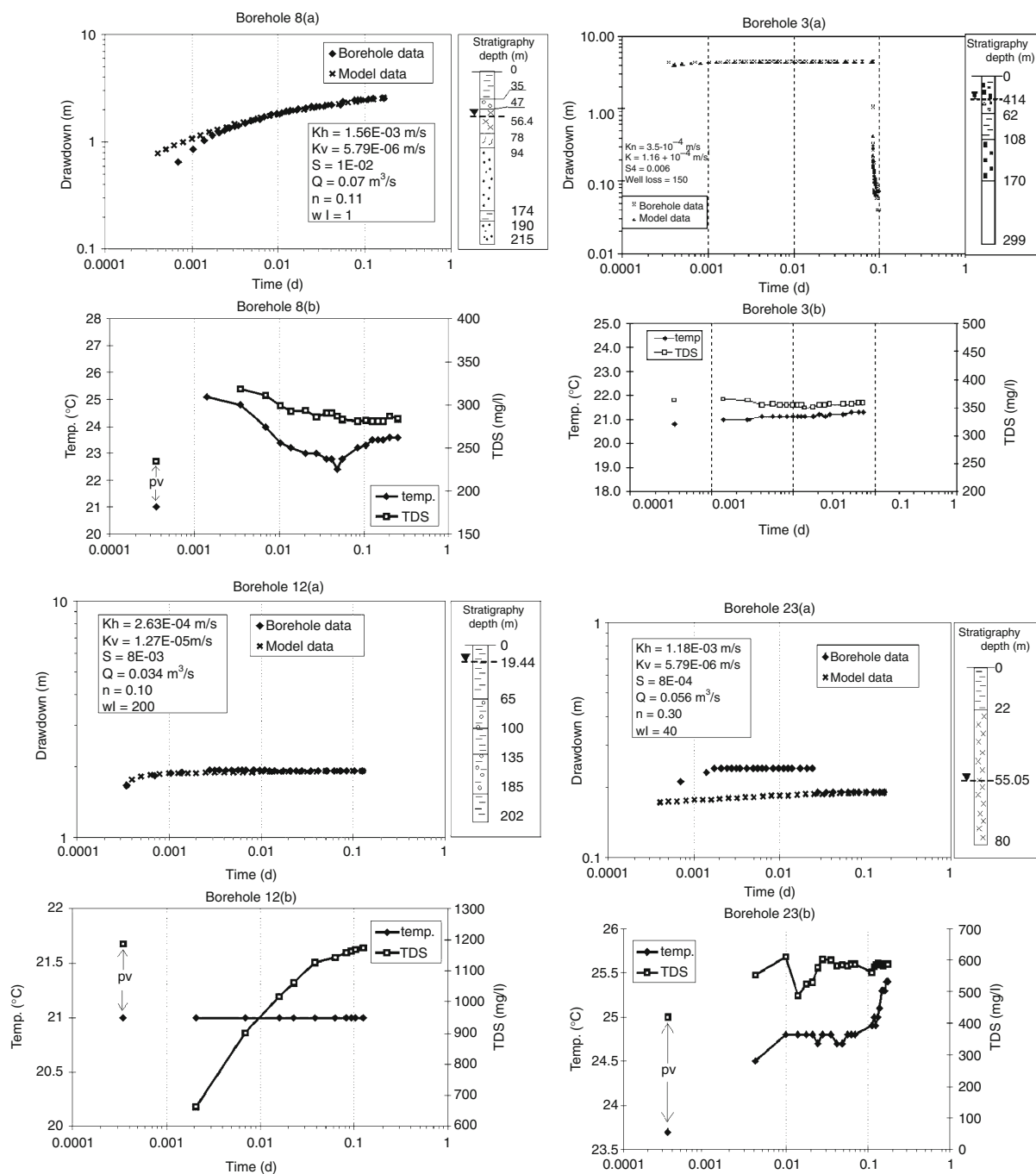
The overall response of each of these boreholes in terms of obtained water quality results from the combined effect of (1) hydrogeological characteristics of affected aquifer environment (i.e., lack or presence of confined conditions, geological structure, Kv/Kh ratio), (2) discharge yield, (3) effect of drawdown on the local

hydraulic potential (i.e., relative thickness among flow systems of different densities), (4) the path length to travel by any flow system to the level of extraction in the particular borehole, (5) borehole construction design as related to the stratigraphic units, and (6) position of the water table. All factors should be further studied to better plan actions to improve the quality of obtained water and to assess possible borehole techniques to better manage flow control. Suggestions on techniques to be applied are proposed in Ref. [8].

As suggested in Fig. 3, the response of borehole 23 is implied to be general to the prevailing long-term extraction yield. The increase over time of groundwater extraction in the basin of Mexico has reached more than 50 m³/s, with individual boreholes extracting an average of 60 L/s. The effect of the partial penetration of production boreholes is that of radial flow in 3-D; relative shallow extraction level in the Mexico City plain results in ascending flow from a depth beyond current extraction levels. This can be interpreted as confirming the presence of the inflow of water of contrasting chemical quality as well as of higher temperature (Fig. 2). This interpretation appears sound when considering the graph of Fig. 4, where mNa/Cl–log Cl suggests that extracted shallow water (local flow) is evolving to water representing the former thermal spring (end-member) of Peñón de los Baños. The inducement of regional flow with different quality into the extraction level augments the concentration of obtained salinity in the water. The vertical components of ascending flow are important in observed water quality changes. Understanding water quality evolution could enhance better control processes of inducement of deep flowing water and aid in predicting where waters of more desirable quality might be obtained as suggested in Ref. [6].

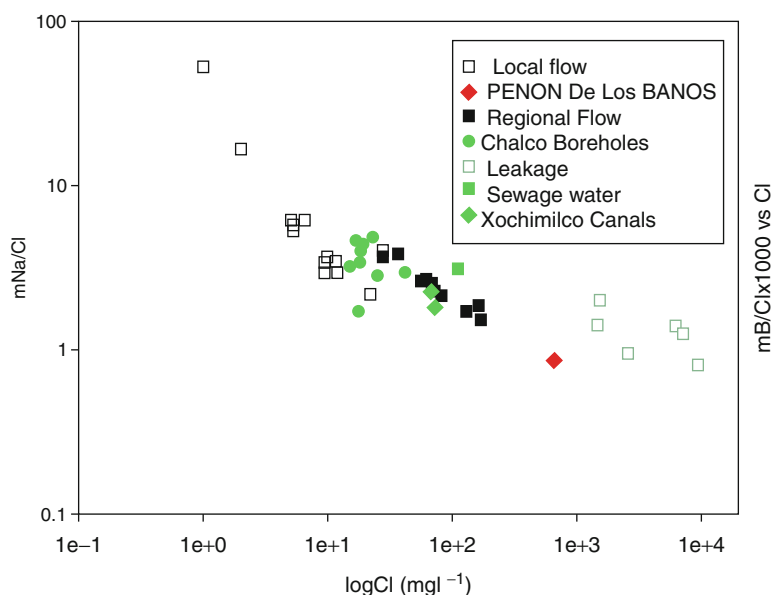
Vertical Flow Control to Prevent Water Quality Deterioration

The identification and understanding of extracted different quality groundwater from contrasting flow systems entering a pumping borehole is one tool for possible control of obtained water quality [8]. These authors interpreted resulting field tests carried out in partially penetrating boreholes tapping a > 2,000-m fractured volcanic (rhyolitic) rock sequence and



Groundwater Salinity Due to Urban Growth. Figure 3

Different responses to groundwater extraction: case 1 (borehole 3) confined conditions; case 2 (borehole 8) semiconfined conditions; case 3 (borehole 23) ascending groundwater flow; case 4 (borehole 12) leakage effect



Groundwater Salinity Due to Urban Growth. Figure 4

Relation of the content of mNa/Cl and Cl for selected boreholes tapping local flows and water mixture representing local and regional end-member of former spring of Peñón de los Baños

proposed a technique to induce the production of the required quality water to a particular extraction borehole. The procedure involves the definition of particular response of flow systems' mixture under prevailing local extraction and hydrogeological variables, such as vertical and horizontal hydraulic conductivities, porosity, hydraulic head response between cold and warm flows, and rate and time of extraction. The response is basically defined through long-term aquifer tests (pumping tests) that include measurement of standard drawdown-time ($s-t$) response plus electrical conductivity (EC) and temperature at the various t intervals of the duration of the test. Additional collection of water samples for chemical and isotopic analyses complements obtained results.

Available information on vertical flow in volcanic terrain with large thickness [8] is important in assessing the factors affecting groundwater flow distribution and obtained water quality in the vicinity of the extraction borehole. An extraction yield will present a particular water quality, depending on the design construction of the borehole, lithology, affected flow systems and chemical components, and temperature of the induced flow systems. Obtaining a specific water quality is diverse and difficult to implement; this

difficulty may be overcome by understanding the control of obtained water quality related to tests carried out in each particular borehole. Methods include an initial step-drawdown test, followed by a constant extraction-rate aquifer test, both in which temperature and EC measurements may be conducted. The response of the borehole could be further defined and understood when full sampling and water analysis is carried out to define major, minor, and trace elements as well as stable isotopes to differentiate the various components of inflowing groundwater flow systems.

Future Directions

Understanding and defining the groundwater system before planning for extraction is a basic rule that requires further attention, especially when practicing in Mexico as well as other countries. The perception of groundwater behavior achieved in the second half of the twentieth century was obtained mainly on function in shallow, thin aquifer units. This acquired knowledge provides a basis to define groundwater response and propose agreeable solutions for those aquifers.

Such powerful techniques have been applied in growing cities as Mexico City, where the increasing need for

water has produced changes in the flow system regime. The continuous development of an interdisciplinary approach as the Flow System Theory [25] is useful for defining the problem and acquiring solutions to the diversity of parameters involved in extracting water of the desired quality. The application of this theory is a powerful tool as it provides answers that are in agreement among the diverse disciplines involved.

Increase in population and economic activities in large cities relies on developing local groundwater flow systems that are often hydraulically linked to regional flow systems when the lithological sequence has a thickness greater than 1,000 m. An interdisciplinary understanding must be reached where groundwater flows of different chemical compositions and temperatures travel in a stratified mode. Analysis by 3-D modeling is useful for clearly defining a problem and proposing adequate solutions when potential impacts on the development of groundwater in thick aquifer units may induce regional flow. Usually, current analysis carried out is made solely on the horizontal scale; however, vertical components of flow need to be acknowledged. Aquifer tests, as well as other common hydrogeological tools, may define groundwater response and its functioning by incorporating the observation of the quality and temperature of extracted groundwater.

Changes in the obtained groundwater quality due to an increase of extraction rate are reflected by the chemistry and temperature of water flowing from a different level to that of the aquifer unit. Water quality changes could become key factors for density variation identification during water extraction, indicating the presence of semiconfining conditions through the behavior of drawdown with time ($s-t$) to that of inflow of water from regional systems at depth.

The growth of cities located on thick aquifer units (i.e., >1,000 m), accompanied with increasing water needs, makes the identification of local, intermediate, and regional groundwater flow systems essential. Two main issues arise in the analysis of thick regional aquifers in Central México: (1) the position of basement rock, which in turn implies the coverage of the aquifer unit involved, and (2) effects due to groundwater chemistry change, such as water density variations, affecting actual hydraulic heads and flow direction. These variables become more important in thick aquifer units.

Bibliography

Primary Literature

1. AIC (1995) El Agua y la Ciudad de México. Academia de la Investigación Científica, Academia Nacional de Ingeniería, Academia Nacional de Medicina, National Academy of Sciences (through the National Research Council), 364 pp
2. Ángeles-Serrano G, Perevochtchicova M, Carrillo-Rivera JJ (2008) Posibles controles hidrogeológicos de impacto ambiental por la extracción de agua subterránea en Xochimilco, México. *J Lat Am Geogr* 7(1):39–56
3. Bouwer H (1978) Groundwater hydrology, McGraw-Hill series in water resources and environmental engineering. McGraw-Hill, Sydney, 480pp
4. Carrillo-Rivera JJ (1998) Monitoring of exploited aquifers resulting in subsidence, example: Mexico City. In: Van Lanen HAJ (ed) Monitoring for groundwater management in (semi-) arid regions, vol 57, Studies and reports in hydrology. UNESCO, Paris, pp 151–165
5. Carrillo-Rivera JJ, Angeles-Serrano G, Hernández GG, Hergt T (2002) Estudio de hidrología subterránea sobre el área de Xochimilco, Distrito Federal. In: Programa rector de manejo del área natural protegida, ejidos de Xochimilco y San Gregorio Atlapulco. Secretaría del Medio Ambiente del Gobierno del Distrito Federal, Distrito Federal
6. Carrillo-Rivera JJ, Cardona A, Edmunds WM (2002) Use of abstraction regime and knowledge of hydrogeological conditions to control high-fluoride concentration in abstracted groundwater: San Luis Potosí basin, Mexico. *J Hydrol* 261:24–47
7. Carrillo-Rivera JJ, Cardona A, Huizar-Alvarez R, Graniel E (2008) Response of the interaction between groundwater and other components of the environment in Mexico. *Environ Geol* 2:303–319
8. Carrillo-Rivera JJ, Cardona A (2008) Groundwater flow system response in thick aquifer units: theory and practice in Mexico. Selected Papers, XXXIII-IAH internacional congress, Zacatecas, México, vol 12. Asociación Internacional de Hidrogeólogos, Editorial Balkema, Taylor & Francis, Leiden, pp 25–46
9. CAVM (Comisión De Aguas Del Valle De México) (1966) Datos del Valle de México, Período 1959–1963. *Bol Mec Suelos* 4
10. Edmunds WM, Carrillo-Rivera JJ, Cardona A (2002) Geochemical evolution of groundwater beneath Mexico City. *J Hydrol* 258:1–24
11. De Cserna Z, De la Fuente DM, Palacios NM, Triay L, Mitre SLM, Mota PR (1988) Estructura geológica, gravimétrica, sismicidad y relaciones neotectónicas regionales de la Cuenca de México, Institute of Geology, UNAM, México. *Boletín* 104:71pp (4maps)
12. Foster S, Garduño H, Evans R, Olson D, Yuan T, Zhang W, Han Z (2004) Quaternary aquifer of the North China Plain – assessing and achieving groundwater resource sustainability. *Hydrogeol J* 12:81–93
13. Freeze A, Cherry J (1979) Groundwater. Prentice, London, 490pp

14. Graniel-Castro D, Morris LB, Carrillo-Rivera JJ (1999) Effects of urbanization on groundwater resources of Mérida, Yucatan. *Environ Geol* 37(4):303–312
15. Herrera I, Alberro J, León JL, Chen B (1974) Análisis de los asentamientos para construcción de los lagos del Plan Texcoco. UNAM, Instituto de Ingeniería, Internal Report No. 340
16. Huizar-Alvarez R, Carrillo-Rivera JJ, Angeles-Serrano G, Hergt T, Cardona A (2004) Chemical response to groundwater extraction southeast of México City. *Hydrogeol J* 12:436–450
17. Llamas MR (1999) La inserción de las aguas subterráneas en los sistemas de gestión integrada. *Bol Geol Min* 110(4):253–370
18. Marsal RJ, Graue R (1969) The subsoil of lake Texcoco. In: Carrillo N (ed) The subsidence of Mexico City and Lake Texcoco project. Secr. de Hacienda y Crédito Publico, México, pp 167–203
19. Mesri G, Rokhsar A, Bohor BF (1975) Composition and compressibility of typical samples of Mexico City clays. *Geotechnique* 25:527–554
20. Ortega GA, Cherry JA, Rudolph DL (1993) Large-scale aquitard consolidation near Mexico City. *Ground Water* 31:708–718
21. Rivera A, Ledoux E (1991) Non-linear modelling of groundwater flow and total subsidence in the Mexico City aquifer-aquitard system. In: Proceedings of the fourth international symposium on land subsidence, Houston, May 1991, IAHS Publication No. 200, pp 45–58
22. Rudolph DL, Cherry JA, Farvolden RN (1991) Field investigations and solute transport in a lacustrine aquitard near Mexico City. *Water Resour Res* 27:2187–2201
23. Theis CV (1935) The relation between the lowering of the piezometric surface and rate and duration of discharge of a well using ground water storage. *Trans Am Geophys Union* 16:519–524
24. Tóth J (1995) Hydraulic continuity in large sedimentary basins. *Hydrogeol J* 3:4–16
25. Tóth J (1998) Groundwater as a geological agent: an overview of the causes, processes, and manifestations. *Hydrogeol J* 7:1–14
26. Tóth J (2008) From the artesian paradigm to basin hydraulics. Eötvös Loránd University. Institute of Geography and Earth Sciences, Budapest
27. Vázquez-Sánchez E (1995) Hidrogeología del Acuífero de la Ciudad de México. M.Sc. thesis in groundwater. UACPyP, Posgrado en Geofísica, Instituto de Geofísica, UNAM
28. Vázquez-Sánchez E, Jaimes-Palomera R (1989) Geología de la Cuenca de México. *Geofis Int* 28(2):133–190
- congress, Zacatecas, México, vol 12. Asociación Internacional Hidrogeólogos, Ed Balkema, Taylor & Francis, Leiden, pp 85–104
- Carrillo-Rivera JJ (1988) The Sierra Madre occidental. In: The geology of North America, hydrogeology, vol 0–2. The Geological Society of North America, Boulder, pp 87–88
- Carrillo-Rivera JJ (1993) The hydrogeology of the San Luis Potosi area, Mexico. *Ground Water J Assoc Ground Water Sci Eng (Abstracts Students Section, Marzo–Abril p 330)*
- Carrillo-Rivera JJ (1998) Monitoring of exploited aquifers resulting in subsidence, example: Mexico City. In: Monitoring for groundwater management in (semi-)arid regions, vol 57, Studies and reports in hydrology. UNESCO, Paris, pp 151–166
- Carrillo-Rivera JJ (2000) Application of the groundwater-balance equation to indicate interbasin and vertical flow in two semi-arid drainage basins, Mexico. *Hydrogeol J* 8(5):503–520
- Carrillo-Rivera JJ (2002) Aquifer evaluation. In: Moore JE (ed) Field hydrogeology. Lewis, Boca Raton, pp 79–87 (Chap 6). ISBN 1-56670-587-8
- Carrillo-Rivera JJ (2003) Lack of a conceptual system view of groundwater resources in Mexico, editor's message. *Hydrogeol J* 11(5):519–520
- Carrillo-Rivera JJ (2004) Congreso conjunto: XXXIII asociación internacional de hidrogeólogos y 7º asociación Latinoamericana de hidrología subterránea para el desarrollo, Zacatecas, Zac. Méx. *Rev Invest Geogr* 55:175–177
- Carrillo-Rivera JJ, Adrián Ortega G (2008) Special number of environmental geology. In: XXXIII international hydrogeologic congress on groundwater flow understanding from local to regional scales, vol 55. Springer, Berlin, pp 235–464. ISSN 0943-0105
- Carrillo-Rivera JJ, Cardona A (2002) Capítulo 14. In: Michael Price (ed) Agua subterránea. Chapman and Hall. Published by Limusa, Noriega Editores, México
- Carrillo-Rivera JJ, Cardona A (2008) Groundwater flow system response in thick aquifer units: theory and practice in Mexico. Selected Papers, XXXIII-IAH internacional congress, Zacatecas, México, vol 12. Asociación Internacional de Hidrogeólogos. Editorial Balkema, Taylor & Francis, Leiden, pp 25–46
- Carrillo-Rivera JJ, Ortega Guerrero MA (eds) (2008) Groundwater flow understanding: from local to regional scales. Selected Papers, XXXIII-IAH internacional congress, Zacatecas, México, vol 12. Asociación Internacional de Hidrogeólogos. Editorial Balkema, Taylor & Francis, Leiden, ISBN: 13 978 0 203 94579 7, p 186
- Carrillo-Rivera JJ, Clark DI, Fritz P (1992) Investigating recharge of shallow and paleo-groundwater in the Villa de Reyes Basin, SLP, Mexico, with environmental isotopes, applied hydrogeology. *Off J Int Assoc Hydrogeol* 4/92:35–48
- Carrillo-Rivera JJ, Cardona A, Moss D (1996) Importance of the vertical component of groundwater flow: a hydrochemical approach in the valley of San Luis Potosí, Mexico. *J Hydrol* 185:23–44
- Carrillo-Rivera JJ, Cardona A, Hergt T (2001) Inducción de agua termal profunda a zonas someras: Aguascalientes, México. *Rev Latinoam Hidrol* 1(1):41–53

Books and Reviews

- Cardona A, Carrillo-Rivera JJ, Huizar-Alvarez R, Graniel-Castro E (2004) Salinization in coastal aquifers of arid zones: an example from Santo Domingo, Baja California Sur, Mexico. *Environ Geol* 45(3):350–366
- Cardona A, Carrillo-Rivera JJ, Castro-Larragoitia GJ, Graniel-Castro EH (2008) Combined use of indicators to evaluate waste-water contamination to local flow systems in semi-arid regions: San Luis Potosi, Mexico. Selected Papers XXXIII-IAH internacional

- Carrillo-Rivera JJ, Varsányi I, Kovács LÓ, Cardona A (2007) Tracing groundwater flow systems with hydrogeochemistry in contrasting geological environments. *Water Air Soil Pollut* 184:77–103
- Carrillo-Rivera JJ, Cardona A, Edmunds WM (2007) Groundwater flow functioning in arid zones with thick volcanic aquifer units: North-Central Mexico. In: International symposium on advances in isotope hydrology and its role in sustainable water resources management (HIS-OIEA) Proc, 21–25 May 2007, vol 1, Vienna, pp 199–211
- Facundo-Castillo JR, Carrillo-Rivera JJ, Antigüedad-Auzmendi I, González Hernández P, Leláes-Díaz R, Hernández-Díaz R, Cáceres-Govea D, Hernández-Santana JR, Suárez-Muñoz M, Melán-Rodríguez C, Rodríguez-Piña M (2008) Chemical and geological control of spring water in eastern Guaniguanico mountain range, Pinar del Rio, Cuba. *Environ Geol* 55(2):247–267
- Fagundo J, Carrillo-Rivera JJ, Antigüedad I, González P, Peláez R, Suárez M, Melián C, Hernández R, Cáceres D (2005) Caracterización hidrogeoquímica del sistema de flujo local-regional de la Sierra del Rosario (Cuba). *Rev Latinoam Hidrol* 5:75–90
- Gutiérrez de MacGregor MT, Carrillo-Rivera JJ, Valdez-Quijada R (1997) Impact of Mexican and USA policies in urban growth and natural resources in the northern border of Mexico, vol 15, Latin American Studies, Nihon-Burajiru Chou Kyokai, Tokio, pp 49–62
- Huizar-Alvarez R, Hernández-García G, Carrillo-Rivera JJ (2009) Simulation of the effects from the groundwater flow on the hydrological balance of the Tecocomulco Lagoon, Central Mexico. *Open Environ Sci J* 3:1–13
- Huizar-Álvarez R, Hernández GG, Carrillo-Martínez M, Carrillo-Rivera JJ, Hergt T, Ángeles-Serrano G (2003) Geologic structure and groundwater flow in the Pachuca–Zumpango sub-basin, central Mexico. *Environ Geol* 43:385–399
- Martínez SE, Carrillo-Rivera JJ (2006) Socio-economic constraints of groundwater in Capital La Rioja, Argentina. *Environ Geol* 49(6):875–886. doi:10.1007/s00254-006-0183-7
- Segovia N, Taméz E, Peña P, Carrillo-Rivera JJ, Acosta E, Armienta MA, Iturbe JL (1999) Groundwater flow system in the valley of Toluca, Mexico: an assay of natural radionuclide specific activities. *Appl Radiat Isot* 50:589–598
- Van-Lanen AJH, Carrillo-Rivera JJ (1998) Framework for groundwater monitoring in (semi-) arid regions. In: Monitoring for groundwater management in (semi-) arid regions, vol 57, Studies and reports in hydrology. UNESCO, Paris, pp 7–20